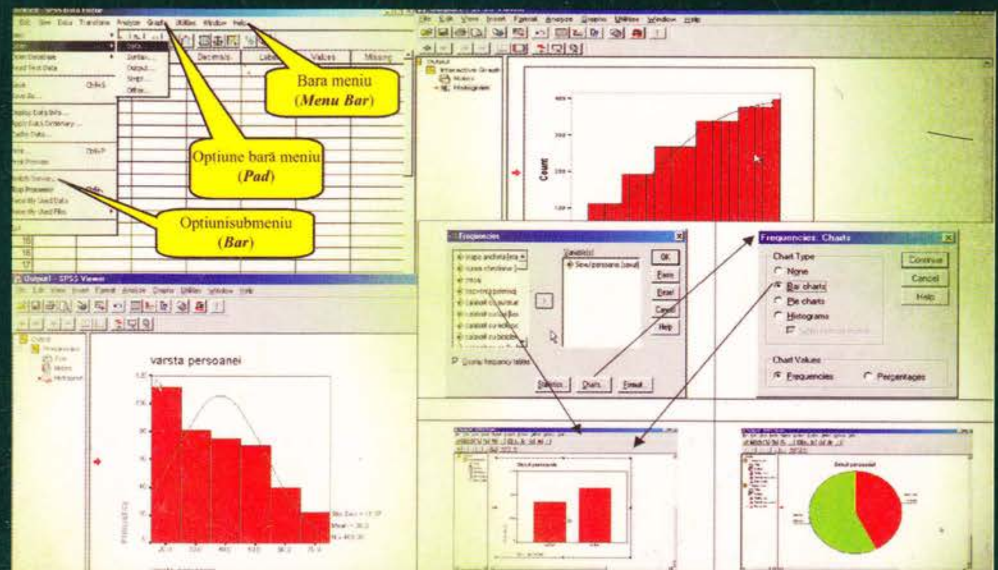
Ştiinţe economice

Elisabeta Jaba, Ana Grama

Analiza statistică cu SPSS sub Windows



POLIROM

Științe economice

Elisabeta Jaba, Ana Grama

Analiza statistică cu SPSS sub Windows

Cartea oferă celor interesați un instrument modern de calcul și analiză a datelor statistice, fiind organizată în acord cu programa analitică folosită în cursurile universitare de Statistică descriptivă și Statistică inferențială. Autoarele prezintă versiunea 10 a SPSS complet integrată în sistemul Windows, punând accent pe folosirea programului SPSS pentru aplicarea diferitelor metode statistice de prelucrare a datelor, în special din economie, sociologie, medicină, și pe interpretarea, în condiții de incertitudine, a rezultatelor statistice din output-ul SPSS.

- Introducere în SPSS
- Elemente conceptuale și metodologice de statistică
- Reprezentarea grafică a unei distribuții în SPSS
- Parametrii unei distribuții statistice
- Distribuția normală
- Estimarea parametrilor unei populații
- Testarea ipotezelor statistice
- Analiza de corelație și regresie

Collegium

ISBN 973-681-609-5





Editura POLIROM
www.polirom.ro

Elisabeta Jaba (n. 1946) este profesor doctor la Universitatea „Al.I. Cuza" din Iaşi, Facultatea de Economie şi Administrarea Afacerilor, şeful Catedrei de Statistică şi prognoză, conducător de doctorat în specializarea Statistică economică. A absolvit Facultatea de Ştiinţe Economice din Iaşi în 1969 şi a obţinut titlul ştiinţific de doctor în economie, specializarea Statistică economică, în 1979. A efectuat mai multe stagii de perfecţionare şi schimburi de experienţă la universităţi din Franţa (în Gestiunea resurselor umane, Anchete prin sondaj, 1992, 1997, 2000), Spania (Statistică şi econometrie, 1992), Anglia (Anchete prin sondaj, 2002), Republica Moldova (1997-2004), Italia (2003). A iniţiat şi coordonat programe SOCRATES şi ERASMUS, în cooperare cu universităţi din Franţa, Italia, Republica Moldova. A publicat, ca unic autor sau coautor, 19 cărţi şi cursuri universitare, precum şi 87 de articole ştiinţifice, în ţară şi în străinătate (China, Franţa, Belgia, Italia). În 2002 a primit Premiul pentru cel mai bun manual al anului în domeniul ştiinţei economice fundamentale, conferit de AGER (Asociaţia Generală a Economiştilor din România), pentru cartea *Statistica*, ediţia a III-a (Editura Economică, Bucureşti, 2002). Este titularul cursurilor de Statistică, Teoria şi practica sondajului statistic, Software statistic. Face parte din Colegiul ştiinţific al *Revistei Române de Statistică*; expert evaluator al CNCSIS. Este preşedintele Societăţii Române de Statistică, filiala Iaşi, şi membru al mai multor societăţi profesionale şi ştiinţifice naţionale şi internaţionale de prestigiu, printre care AGER şi Societatea Franceză de Statistică.

Ana Grama (n. 1953) este profesor universitar la Universitatea „Al.I. Cuza" din Iaşi, Facultatea de Economie şi Administrarea Afacerilor, Catedra de Informatică economică. Din 1998 este doctor în economie, specializarea Informatică economică. A efectuat stagii de perfecţionare în Franţa (Toulouse) şi Italia (Roma). Este autor sau coautor a 27 de cărţi, cursuri universitare şi lucrări practice şi a publicat 50 de articole şi studii din domeniul Informaticii economice. Este titular al cursurilor de Introducere în informatica economică şi Medii de programare.

Referenţi ştiinţifici:

Prof.univ.dr. Alecsandru-Puiu Tacu

Prof.univ.dr. Dumitru Oprea

www.polirom.ro

Editura POLIROM
Iaşi, B-dul Carol I nr. 4, P.O. BOX 266, 700506
Bucureşti, B-dul I.C. Brătianu nr. 6, et. 7, ap. 33, O.P. 37; P.O. BOX 1-728, 030174

Descrierea CIP a Bibliotecii Naţionale a României:

JABA, ELISABETA

Analiza statistică cu SPSS sub Windows / Elisabeta Jaba, Ana Grama - Iaşi: Polirom, 2004

272 p.; 24 cm (Collegium. Ştiinţe economice)

ISBN: 973-681-609-5

I. Grama, Ana

004:311

Printed in ROMANIA

Elisabeta Jaba, Ana Grama

Analiza statistică cu SPSS sub Windows

POLIROM
2004

CUPRINS

Capitolul 3
Pregătirea, sistematizarea şi prezentarea datelor în SPSS

Capitolul 4
Reprezentarea grafică a unei distribuţii în SPSS

Capitolul 5
Parametrii unei distribuții statistice

Capitolul 6
Distribuția normală

Cuvânt înainte

Statistica studiază fenomene de masă, produse sub semnul incertitudinii. Fundamentarea deciziilor cu privire la astfel de fenomene, în condiţiile ritmului trepidant al vieţii contemporane, necesită tot mai multă informaţie, de o calitate tot mai bună, obţinută într-un timp cât mai scurt. Ca răspuns la o asemenea cerere are loc perfecţionarea metodelor şi instrumentelor de obţinere a datelor statistice, precum şi a instrumentelor de calcul. Au fost elaborate programe speciale de prelucrare şi analiză statistică, printre care *SPSS* este unul dintre cele mai actuale, mai performante, mai cunoscute şi mai larg răspândite.

Elaborarea lucrării de faţă a pornit de la o astfel de necesitate. Scopul propus este de a-i familiariza pe cei interesaţi cu un instrument modern de calcul şi analiză a datelor statistice, precum şi cu modul de interpretare statistică a rezultatelor.

Lucrarea reprezintă primul volum din ciclul *ANALIZA STATISTICĂ CU SPSS SUB WINDOWS*: *Vol. I – Statistica descriptivă şi inferenţială; Vol. II – Statistică avansată; Vol. III – Analiza seriilor de timp.*

Informaţia din această carte este organizată în acord cu programa analitică folosită în cursurile universitare de Statistică descriptivă şi Statistică inferenţială.

Pentru obţinerea informaţiei statistice, în acest volum este întreprins un demers care, pe de o parte, se bazează pe conceptele şi metodele clasice ale statisticii şi, pe de altă parte, este asistat de calculator prin programul SPSS. Lucrarea este structurată astfel încât să permită înţelegerea şi însuşirea conceptelor fundamentale ale statisticii şi utilizarea metodelor statistice, folosind programul SPSS pentru rezolvarea problemelor de prelucrare şi analiză statistică.

Dacă până acum se punea accent pe însuşirea conceptelor şi tehnicilor de calcul manual al indicatorilor statistici, în această lucrare accentul se mută de la calcul la interpretare. Locul timpului consumat cu prelucrarea manuală a datelor este cedat, prin exploatarea programului SPSS, analizei, simulării şi interpretării rezultatelor. Acest lucru este important pentru obţinerea în timp util a unei informaţii statistice de calitate, dintr-o bază de date specifică fenomenelor de masă, cum ar fi datele rezultate dintr-o anchetă statistică sau dintr-un experiment, realizate asupra unor colectivităţi statistice, pentru care trebuie fundamentate, în timp real, decizii de politică economică, socială şi de altă natură.

Caracteristica majoră care face din lucrare un instrument pentru profesionişti este prezentarea analizei statistice cu SPSS sub Windows.

Versiunea 10 a SPSS folosită este complet integrată în mediul WINDOWS; dezvoltarea instrumentului informatic favorizează utilizarea acestui program în analiza datelor statistice.

Cartea se doreşte a fi un suport didactic pentru statistică, autorizat, complet şi actualizat, pentru studenţii de la facultăţile de economie, sociologie, medicină, farmacie şi nu numai. Oferă, de asemenea, elemente care o fac utilă practicienilor din astfel de domenii. Se pune accent *pe folosirea programului SPSS pentru aplicarea diferitelor metode statistice de prelucrare a datelor*, în special din economie, sociologie, medicină, şi *pe interpretarea, în condiţii de incertitudine, a rezultatelor statistice din output-ul SPSS.*

Într-o lume a relativităţii, ce este mai bine de ales: *o certitudine iluzorie* (deoarece totul este relativ) *sau o incertitudine măsurabilă* (probabilă)? *Răspunsul* pe care îl dă lucrarea este unul *statistic.*

Dedicăm cartea cititorilor de toate vârstele, *virusaţi* de freamătul neliniştii generate de pasiunea de a găsi calea *lucrului bine făcut.*

Dacă sunteţi interesaţi, vă invităm să parcurgeţi paginile lucrării *Analiza statistică cu SPSS sub Windows*, elaborată în condiţiile unei colaborări deosebite între un statistician – profesor universitar dr. *Elisabeta Jaba* – şi un informatician – profesor universitar dr. *Ana Grama.*

Autoarele, conştiente că orice lucru este perfectibil, mulţumesc anticipat cititorilor pentru bunăvoinţa de a sesiza neîmplinirile şi de a le transmite sugestii pentru îmbunătăţirea ediţiei actuale.

Autoarele
august 2003, Iaşi

CAPITOLUL 1

INTRODUCERE ÎN SPSS

- **Componente şi caracteristici**
- **Sesiunea de lucru SPSS**
- **Ferestrele SPSS**
- **Gestiunea fişierelor SPSS**

SPSS (*Statistical Package for the Social Sciences*) este unul dintre cele mai puternice şi utilizate programe statistice[1]. Acest pachet integrat asigură acoperirea procedeelor specifice din *Statistica descriptivă*, *Statistica inferenţială* şi *Analiza datelor*. Programul a devenit deosebit de atractiv pentru utilizatori deoarece permite tratarea datelor statistice fără a impune cunoaşterea formulelor de calcul, îmbinând posibilităţile de prelucrare statistică cu facilităţile oferite de programele de calcul tabelar (Excel, Lotus, Quattro Pro) pentru condensarea datelor în tabele şi reprezentarea lor grafică.

Programul este un produs al firmei SPSS Inc., care s-a impus în domeniul realizării de software pentru prelucrarea statistică a datelor, în principal prin SPSS şi SYSTAT.

SPSS a fost creat la Universitatea din Stanford, în anii '60, de către doi studenţi, Norman Nie şi Tex Bull, pentru a asigura gestiunea şi analiza datelor statistice în domeniul ştiinţelor sociale şi al psihologiei. Ulterior, utilizarea programului s-a extins spre economie, medicină etc. În acelaşi timp, evoluţiile din domeniul calculatoarelor au marcat şi dezvoltarea SPSS, prin apariţia imediată a unor noi versiuni.

Începând cu versiunea 7, realizată în 1995, SPSS a devenit un produs pentru Windows, ajungându-se astăzi la versiunea 12.

După 30 de ani de la crearea sa, SPSS este folosit în peste 2.500 de universităţi şi instituţii de învăţământ superior şi în peste 250.000 de instituţii din diverse sectoare (administraţie, educaţie, lumea afacerilor etc.).

1.1 Componente şi caracteristici

1.1.1 Modulele SPSS

La ora actuală, SPSS este realizat sub formă modulară, fiecare utilizator putându-şi achiziţiona doar acele componente care îi sunt necesare. Cele mai „comercializate" module sunt: *Base module*, *Professional Statistics*, *Advanced Statistics*, *Tables*, *Exact Tests*, *CHIAD* şi *Categories*.

Modulul de bază – *Base module* – permite gestionarea datelor şi fişierelor, transformarea datelor, precum şi prelucrarea statistică a acestora prin:

- calculul frecvenţelor, al indicatorilor tendinţei centrale, dispersiei şi formei unei distribuţii;

1. Din această categorie mai fac parte: STATISTICA, SAS, SYSTAT, S-PLUS, R-project etc.

- calculul măsurilor de asociere şi testarea independenţei probabilistice pentru date incluse în tabelele de contingenţă;
- compararea mediilor, proporţiilor şi dispersiilor eşantioanelor;
- analiza unifactorială a varianţei;
- calculul coeficienţilor de corelaţie Pearson, Kendall, Spearman;
- analiza de regresie liniară;
- teste neparametrice.

De asemenea, acest modul permite şi reprezentarea grafică a datelor sub formă de histograme, diagrame de structură, nor de puncte etc.

Modulul *Professional Statistics* include proceduri pentru cercetarea relaţiilor dintre variabile, folosind ca metode:
- analiza de discriminant;
- analiza factorială;
- analiza de clusteri;
- scalarea multidimensională;
- regresia ponderată;
- analiza fidelităţii.

Modulul *Advanced Statistics* permite efectuarea unor prelucrări statistice complicate, apelând la următoarele metode:
- analiza de regresie logistică;
- diverse extinderi ale analizei unifactoriale a varianţei ANOVA;
- analiza varianţei multifactorială MANOVA;
- analiza logliniară;
- analiza de regresie neliniară;
- analiza probit şi logit;
- analiza duratei de viaţă;
- analiza de supravieţuire Kaplan-Meier;
- modelul liniar general.

Modulul *Tables* permite condensarea datelor în tabele cu una, două sau trei dimensiuni, fiecare dimensiune fiind definită printr-o variabilă sau printr-un grup de variabile. Pe lângă valorile variabilelor, tabelele pot conţine frecvenţe şi valori ale unor indicatori statistici (medie, abatere standard etc.).

Modulul *Exact Tests* determină nivelul de semnificaţie (valorile *p* sau *Sig.*) pentru:
- teste neparametrice aplicate pe un eşantion, pe două eşantioane, independente sau perechi, şi pe k eşantioane dependente sau independente;
- teste aplicate tabelelor de contingenţă 2×2 şi $r \times c$;

- teste de semnificație pentru coeficienții de corelație Pearson și Spearman;
- teste referitoare la relațiile dintre variabile măsurate pe scală nominală sau pe scală ordinală.

Modulul *CHIAD* (Chi-squared Automatic Interaction Detector) aplică algoritmi de segmentare pentru împărțirea unei populații în grupe disjuncte, care diferă între ele în funcție de un criteriu precizat. La fiecare pas al algoritmului, grupele constituite sunt vizualizate sub forma *dendrogramelor*.

Modulul *Categories* este folosit pentru determinarea influenței exercitate de caracteristicile produselor și serviciilor asupra preferinței consumatorilor. Pentru identificarea asemănării sau deosebirii dintre obiecte, acest modul vizualizează prin puncte obiectele analizate.

Modulul *TRENDS* asigură analiza și reprezentarea grafică a seriilor de timp. Este posibilă estimarea coeficienților modelului de trend prin următoarele tehnici:

- procedee de ajustare;
- metode de regresie;
- analiza Box-Jenkins (ARIMA);
- procedee de descompunere sezonieră, pentru determinarea factorilor aditivi și multiplicativi, în cazul seriilor de timp cu caracter sezonier;
- analiza componentei aleatorii.

1.1.2 Caracteristici ale SPSS

Dacă ar fi să caracterizăm acest produs prin acronimul de apelare, *SPSS* s-ar putea evidenția prin:

- *S*oluții pentru probleme complexe;
- *P*rezentarea sugestivă a rezultatelor;
- *S*uplețe în stabilirea condițiilor de prelucrare a datelor prezente într-o mare diversitate;
- *S*implitate în exploatare.

Soluții pentru probleme complexe. Având la dispoziție instrumente specifice metodelor statistice avansate, SPSS permite rezolvarea problemelor oricât de complexe, din diverse domenii, oferind soluții care să asigure o cunoaștere mai bună a fenomenelor cercetate și, implicit, să sprijine procesul de fundamentare a deciziilor.

Prezentarea sugestivă a rezultatelor. Utilizatorul are control deplin asupra tuturor variabilelor prelucrate, stabilind modul de afişare a valorilor din listele de ieşire (lungime, număr de zecimale) şi ce text să fie scris în locul denumirilor variabilelor (atunci când acestea nu sunt destul de sugestive) sau în locul valorilor variabilelor (dacă în fişierul de date s-au introdus coduri). Pentru mai multe detalii, vezi paragraful 3.4.

Listele de rezultate, tabelele şi graficele realizate de SPSS pot fi incluse în rapoarte, aşa cum se prezintă pe ecran, sau într-o formă modificată prin:

- editarea de texte;
- stabilirea caracteristicilor fonturilor/caracterelor (tip, stil, mărime, culoare);
- modificarea desenelor prin deplasarea şi/sau rotirea axelor ori schimbarea tipului de grafic;
- ascunderea unor variabile din tabele;
- reorganizarea informaţiilor în tabele (de exemplu, un tabel de frecvenţe care conţine pe linii răspunsurile la un chestionar, iar pe coloane localitatea de domiciliu şi, în cadrul fiecărei localităţi, sexul clienţilor, poate fi transformat într-un tabel cu numai două coloane, corespunzător sexului, şi cu grupe de linii, câte o grupă pentru fiecare localitate).

Toate aceste operaţii sunt uşor de executat, datorită existenţei a trei editoare: de *text*, de *tabele* şi de *grafice*. Rezultatele prelucrărilor statistice se pot vizualiza prin tabele de diverse formate şi prin multiple tipuri de reprezentări grafice: *histograme*, *diagrame „coloane"* – izolate sau grupate –, *diagrame de structură*, *nor de puncte* – în care punctele corespunzătoare unor grupe diferite sunt colorate diferit –, *diagrame „bare"* – care indică în acelaşi timp media, valorile extreme şi repartiţia valorilor unei variabile pentru valori diferite ale altei variabile (de exemplu, reprezentarea grafică a distribuţiei după vârsta persoanelor, în funcţie de localitatea de domiciliu). În grafice, se depistează rapid valorile „aberante", valori izolate, semnificativ diferite de restul datelor din fişier (*outlier*-i).

Supleţe în stabilirea condiţiilor de prelucrare a datelor. Domeniile diferite în care SPSS îşi găseşte aplicare oferă o mare diversitate a condiţiilor de prelucrare. SPSS permite realizarea oricărei variante de prelucrare, ori de câte ori este nevoie, la nivelul întregii baze de date sau la nivelul unui subansamblu de date selectat.

Dacă un grup de prelucrări se efectuează periodic (de exemplu, dacă interesează situaţia zilnică a vânzărilor pe magazine şi produse), întreaga succesiune de căutări prin meniuri şi de alegeri de opţiuni nu se repetă de

fiecare dată. SPSS poate înregistra într-un fişier de comenzi toate aceste operaţii. Ulterior, fişierul va fi rulat ori de câte ori este necesar. În plus, fişierul poate fi actualizat, în sensul că i se pot adăuga sau şterge comenzi.

Utilizatorul poate alege cazurile care să fie luate în considerare la efectuarea prelucrărilor, formulând condiţii asupra uneia sau mai multor variabile. De asemenea, utilizatorul poate decide modul în care să fie tratate de SPSS cazurile în care valoarea unei variabile nu este cunoscută sau nu prezintă interes pentru cercetare. Ele pot fi sau nu incluse în calcule.

Înainte de efectuarea prelucrărilor statistice, SPSS poate modifica automat datele pe baza unor algoritmi indicaţi de utilizator pentru recodificarea valorilor sau prin aplicarea unor funcţii matematice. De exemplu, într-un fişier în care cazurile sunt reprezentate de diverse mărfuri, iar variabilele de însuşirile acestora, preţurile mărfurilor pot fi schimbate prin adăugarea TVA sau toate mărfurile produse înainte de 2000 pot primi aceeaşi valoare a variabilei „data de fabricaţie", care să semnifice „înainte de 2000".

Simplitate în exploatare. SPSS este un program care poate fi exploatat şi de persoane mai puţin iniţiate în statistică. Meniul *Help* permite accesul la un glosar de termeni care prezintă semnificaţia acestora, în meniuri şi casete de dialog, iar componenta *Tutorial on-line* aduce explicaţii şi exemple care permit orientarea rapidă printre numeroasele prelucrări care pot fi realizate.

Pentru orice noţiune, dintr-o căsuţă de dialog sau chiar dintr-o listă de ieşire, se obţine afişarea unui text explicativ (help/ajutor contextual) dacă se alege opţiunea *What's This?*.

Cât priveşte exploatarea propriu-zisă, SPSS asigură simplitate în manevrarea datelor de intrare. Introducerea şi modificarea datelor este o operaţie simplă datorită existenţei unui editor de tabele de tip *spreadsheet*. Pe ecran este afişat un tabel cu linii şi coloane. Liniile corespund *cazurilor* (subiecţi care răspund unui chestionar sau obiecte observate), iar coloanele conţin *variabilele* (răspunsuri date de subiecţi sau rezultatele unor măsurători ori observaţii). Nu există limitări în privinţa numărului de cazuri sau variabile care pot fi incluse în tabel (fişier). Utilizatorul poate „naviga" prin acest tabel, după dorinţă, analizând valorile existente, schimbând unele date, adăugând sau ştergând cazuri şi variabile. SPSS adaptează automat dimensiunile tabelului, astfel încât să nu se piardă nici o valoare introdusă.

SPSS asigură prelucrarea datelor preluate din registrele de lucru Excel, Lotus 1-2-3, bazele de date dBase sau fişierele de text ASCII. În acelaşi timp, fişierele create în SPSS pot fi exportate în Excel, Lotus 1-2-3 sau fişiere text ASCII.

Rezumând cele de mai sus, se poate aprecia că SPSS este un produs orientat spre utilizatorul-analist şi permite:

- analiza datelor sub multiple aspecte;
- extinderea datelor cercetărilor realizate pe un eşantion, la nivel naţional;
- construirea tabelelor de ieşire în diverse forme, inclusiv cu totaluri şi structuri pe orizontală şi verticală;
- construirea diagramelor sub diferite forme (linii, bare, sectoare etc.);
- crearea prezentărilor şi a rapoartelor;
- utilizarea datelor în regim interactiv şi construirea seturilor de funcţii ale sistemului pentru folosirea lor repetată (automatizarea analizei datelor);
- exploatarea facilităţilor oferite de Internet;
- elaborarea unor programe de introducere şi control al datelor.

1.2 Sesiunea de lucru SPSS

Perioada de timp în care sunt exploatate facilităţile oferite de SPSS poartă numele de *sesiune de lucru*. În acest interval utilizatorul lansează comenzi pentru realizarea anumitor operaţii, iar sistemul afişează rezultate şi/sau mesaje. Dialogul utilizator-calculator este interactiv, interfaţa avantajând chiar şi un utilizator începător, în sensul că ferestrele deschise oferă variantele de lucru, din care se poate alege succesiunea etapelor pe care trebuie să le urmeze în prelucrarea datelor. Acest lucru este posibil pentru că programul citeşte datele şi le transformă la cerere, prin operaţii matematice şi statistice.

1.2.1 Deschiderea şi închiderea unei sesiuni de lucru SPSS

După *instalare*[2], pachetul SPSS, pentru a fi exploatat, poate fi lansat în două moduri:

- folosind pictograma *SPSS* de pe suprafaţa *Desktop* (vezi figura 1.1), dacă anterior a fost creată o scurtătură (*Shortcut*);

2. *Instalarea* este operaţia prin care produsul-program SPSS este încărcat de pe un suport extern (numit *kit de instalare* – CD sau dischetă) pe hard disk-ul sistemului de calcul.



Figura 1.1 *Scurtătura SPSS*

- folosind din bara de task-uri[3] butonul *Start*, din care se selectează succesiv: *Programs* → *SPSS for Windows* (vezi figura 1.2).

Observație! În această lucrare, se utilizează ca sistem de operare *Windows XP*.

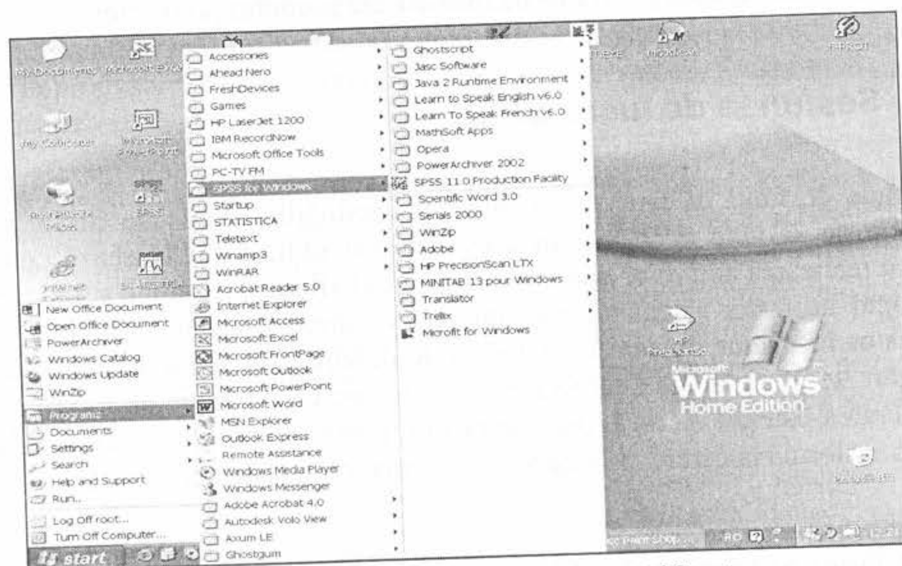


Figura 1.2 *Apelarea* SPSS *din meniul* Start

Închiderea unei sesiuni de lucru *SPSS* se poate realiza prin:

- butonul ☒ din bara de titlu a unei ferestre principale;
- comanda *Exit* din meniul *File*;
- comanda *Close* (sau combinația de taste *Alt* + *F4*) din meniul de control al unei ferestre[4].

3. *Task*-ul reprezintă o operație/sarcină/lucrare sau grup de acțiuni ce formează o unitate logică, din punctul de vedere al sistemului de operare (în cazul de față Windows, care, se știe, este un sistem multitasking).
4. *Meniul de control* este atașat ferestrelor Windows, iar pictograma de activare este plasată în partea stângă a barei de titlu (prima linie dintr-o fereastră Windows).

1.2.2 Interfaţa SPSS

SPSS exploatează o interfaţă de tip *WIMP* (Window, Icon, Mouse, Pulldown), în care elementele de bază sunt ferestrele, pictogramele (icoanele), mouse-ul şi meniurile derulante (*pulldown menus*).

Ferestrele sunt zone/porţiuni de pe ecran tratate ca elemente de sine stătătoare, cu caracteristici proprii, care determină acţiunile ce se pot executa în cadrul lor. O fereastră este afişată ca urmare a lansării unei anumite operaţii.

Ferestrele pot fi *principale* şi *de dialog*. Cele principale sunt subordonate direct operaţiilor declanşate, iar cele de dialog permit utilizatorului să stabilească sau să selecteze condiţiile de derulare a operaţiilor.

Pictogramele se prezintă sub forma unor mici imagini însoţite de un text care sugerează programul, funcţia sau comanda pentru care au fost create. Practic ele sunt scurtături (*shortcut*-uri) pentru programe, comenzi etc. De regulă, ele apar pe suprafaţa *Desktop* sau sunt plasate în barele de instrumente (*Toolbars*) ale ferestrelor principale.

Mouse-ul este dispozitivul periferic de intrare folosit pentru selectarea şi lansarea rapidă a comenzilor şi este aproape indispensabil pentru o interfaţă grafică. Face parte din configuraţia minimă a unui sistem electronic de calcul.

Meniurile reprezintă elemente prin care i se oferă utilizatorului posibilitatea selectării unei anumite opţiuni dintr-o mulţime finită.

Un meniu conţine următoarele elemente (vezi figura 1.3):

- bara meniu (*menu bar*);
- opţiunile barei meniu (*pad*-uri);
- submeniuri (*popup*-uri sau *submenu*-uri);
- opţiunile submeniurilor (*bar*-uri).

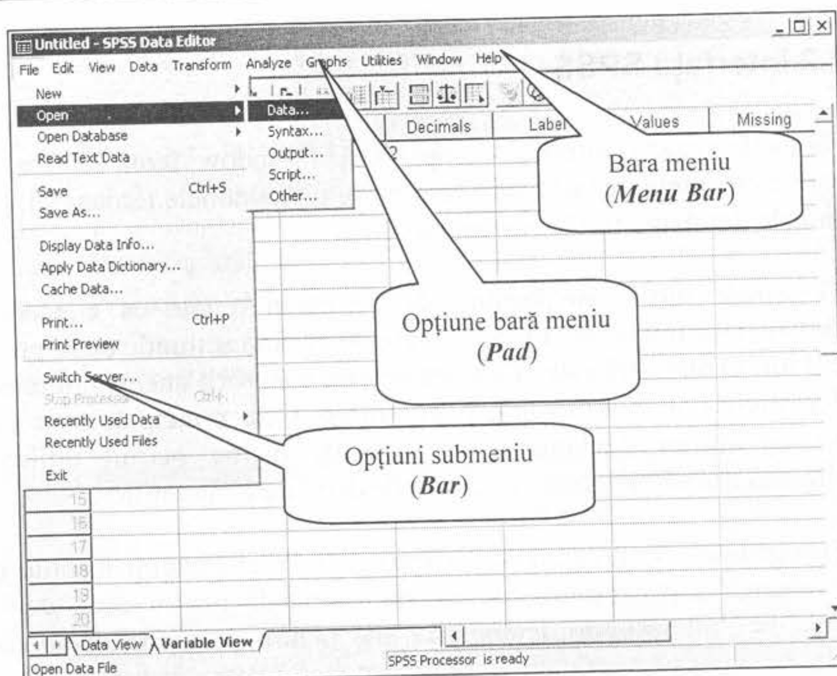


Figura 1.3 *Organizarea meniurilor*

Bara meniu (*menu bar*) este structurată pe orizontală şi asigură organizarea tuturor celorlalte elemente componente ale unui meniu. În această bară sunt plasate *pad*-urile care, prin numele lor, sugerează funcțiile pe care le pot îndeplini opțiunile barei meniu. De regulă, un *pad* are în subordine un submeniu (*popup*) în structura căruia intră opțiuni (*bar*-uri) organizate pe verticală.

1.3 Ferestrele SPSS

SPSS lucrează cu mai multe ferestre diferite, fiecăreia dintre ele fiindu-i asociat un anumit tip de fişier. Dintre acestea, pentru analiza datelor, cele mai frecvent utilizate sunt ferestrele *Data Editor*, *Syntax Editor* şi *Output Viewer*. Pe lângă acestea sunt utilizate şi alte ferestre, specializate în editarea de text, grafice, tabele.

Fereastra de editare (*Data Editor*) se deschide implicit la lansarea SPSS şi este folosită pentru introducerea, modificarea sau ştergerea datelor în format *spreadsheet*. Într-o fereastră de editare poate fi prezentat conținutul unui fişier de date care a fost selectat dintr-o listă de fişiere create anterior (în SPSS,

Excel, Statistică etc.) sau poate fi creată o nouă foaie de date. Această fereastră recunoaşte *fişierele de date* care au extensia implicită *.sav*.

Fereastra de sintaxă (*Syntax Editor*) este folosită pentru a genera programe de comenzi pe care dorim să le executăm asupra datelor (de exemplu, transformarea datelor, calculul unor noi variabile ş.a.). Opţiunile selectate în casetele de dialog sunt afişate în fereastra de sintaxă sub formă de comenzi. Acestei ferestre îi sunt specifice fişierele de tip *.sps*.

Fereastra de rezultate (*Output Viewer*) devine disponibilă automat după ce a fost efectuată o comandă de analiză a datelor. În această fereastră, sunt afişate rezultate statistice, tabele şi grafice care au asociate ferestre distincte.

Fereastra de editarea a rezultatelor (*Text Output Editor*) este folosită pentru modificarea textului rezultat, care nu a fost afişat în tabele pivot.

Fereastra Pivot Table (*Pivot Table Editor*) oferă multiple posibilităţi de modificare a tabelelor pivot: editare text, schimbarea datelor din rânduri şi coloane, adăugarea de culori, crearea unor tabele multidimensionale, ascunderea sau afişarea selectivă a rezultatelor.

Fereastra de editare a graficelor (*Chart Editor*) permite modificarea elementelor unui grafic (axe, scale, diagramă, legendă etc.).

1.3.1 Fereastra Data Editor

În fereastra *Data Editor* sunt afişate datele de lucru. Acestea sunt aranjate în format tabel (*spreadsheet*), care conţine coloane şi linii. La intersecţia acestora sunt celulele (casetele, căsuţele) în care se introduc datele. La un moment dat, este activă (curentă) o singură celulă, cea în care este plasat cursorul. Celula curentă este scoasă în evidenţă printr-un chenar îngroşat. Trecerea de la o celulă la alta se realizează prin clic de mouse în noua celulă, sau de la tastatură cu ajutorul tastelor de control al mouse-ului (tastele săgeţi şi PageUp/ PageDown).

Întotdeauna, coloanele tabelului reprezintă *variabilele* cercetării. De altfel, denumirea coloanelor – *var* – sugerează conţinutul acestora. Liniile tabelului sunt numerotate şi reprezintă *cazurile* (subiecţii sau participanţii la cercetare).

Fereastra *Data Editor* conţine două foi: *Data View*[5] şi *Variable View*[6] (vezi figura 1.4). La un moment dat este activă/vizibilă una singură, şi anume cea în

5. *Data View* este ca un *worksheet* (foaie de calcul) din programul de calcul tabelar *Excel*.
6. *Variable View* este ca o fereastră *Table Designer view* din sistemele de gestiune a bazelor de date *Acces* sau *FoxPro*.

care este plasat cursorul (pointer-ul sistem). Fiecare foaie are, în partea de jos a suprafeţei de lucru, câte o etichetă (*Label*) cu numele ei. Trecerea dintr-o foaie în alta se realizează printr-un clic de mouse de pe eticheta proprie foii respective. La deschiderea editorului de texte, este vizibilă foaia *Data View* care conţine datele brute. În aparenţă, cea de a doua foaie (*Variable View*) este similară cu prima, dar ea conţine nu date, ci informaţii despre variabilele de analizat (nume – *Name*, tip – *Type*, lungime – *Width* etc.).

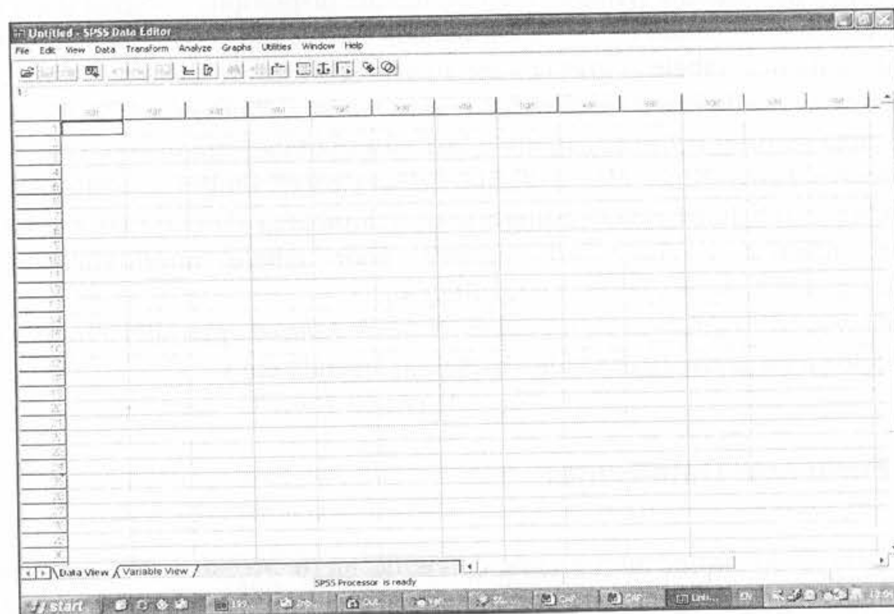


Figura 1.4 *Fereastra* Data Editor *cu foile* Data View *şi* Variable View

Până la versiunea 10 SPSS, fereastra *Data Editor* permitea deschiderea, la un moment dat, a unei singure baze de date (set de date). Începând cu această versiune, pot fi deschise, în acelaşi timp, mai multe ferestre Data Editor, fiecare conţinând o altă bază de date. Activă este însă una singură şi se numeşte bază de date de lucru (*working datasets*). Asupra acesteia sunt executate toate manipulările, funcţiile statistice şi alte proceduri SPSS.

Ca orice fereastră Windows, şi cea a editorului de texte SPSS organizează mai multe meniuri, foarte utile pentru execuţia unor operaţii variate asupra datelor.

1.3.2 Fereastra Syntax Editor

Versiunile mai recente ale SPSS conțin meniuri *pull-down* şi casete de dialog care permit lansarea comenzilor SPSS fără a scrie sintaxa acestora. Tutorialele SPSS se concentrează pe utilizarea casetelor de dialog pentru execuția procedurilor.

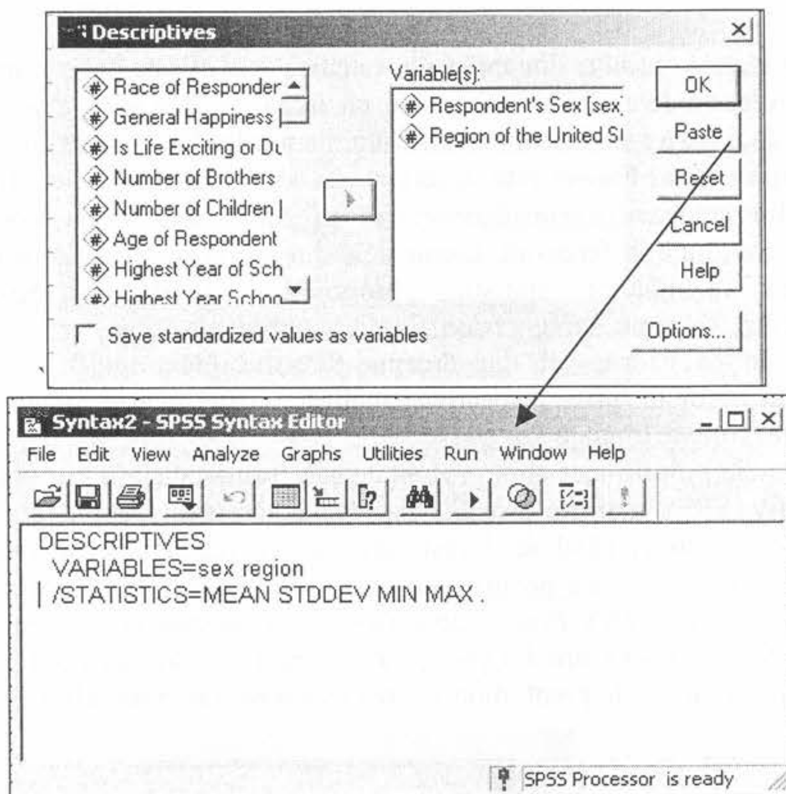


Figura 1.5 *Fereastra SPSS* Syntax Editor

Există şi situații în care casetele de dialog nu pot răspunde tuturor cererilor de prelucrare. Pe de o parte, nu toate procedurile de prelucrare sunt disponibile în casetele de dialog, motiv pentru care se impune utilizarea *Syntax Editor*-ului. Pe de altă parte, există situații în care procedurile nu pot fi salvate ca sintaxă pentru a fi relansate ulterior. Casetele de dialog disponibile în meniurile *pull-down* sunt prevăzute cu câte un buton de comandă *Paste* care are rolul de a „tipări" sintaxa pentru procedura realizată prin mediul oferit de caseta de dialog în fereastra *Syntax Editor* (vezi figura 1.5).

Procedura astfel obţinută poate fi salvată şi ulterior executată, dacă baza de date activă în fereastra *Data Editor* conţine variabile cu acelaşi nume. Fişierul salvat are extensia *.sps*. Salvarea sintaxei este utilă mai ales atunci când aceeaşi analiză trebuie executată şi asupra altei baze de date, dar care conţine aceleaşi variabile.

1.3.3 Fereastra Output Viewer

Toate rezultatele obţinute din analizele statistice sunt afişate în fereastra *Output Viewer*. Această fereastră se aseamănă cu fereastra *Windows Explorer* şi se deschide doar dacă s-au lansat comenzi din meniurile *Analyze* sau *Graphs*.

Fereastra *Output Viewer* este structurată în două cadre/zone (vezi figura 1.6). Cadrul din stânga (*cuprinsul/structura*) prezintă, sub forma unei *schiţe*, obiectele conţinute în fereastră. Elementele din schiţă se referă la titlu, note şi denumirea rezultatelor statistice propriu-zise (Statistică descriptivă – *Descriptives*, Regresie – *Regression*, Grafic – *Graph* etc.).

În al doilea cadru, cel din dreapta ferestrei (*conţinutul*), sunt afişate rezultatele obţinute prin respectiva analiză. Toate aceste obiecte pot fi modificate, copiate, mutate sau şterse.

Rezultatele propriu-zise sunt prezentate sub formă tabelară sau grafică. Din acest motiv, SPSS mai are asociate încă două ferestre, *Pivot Table* şi *Chart Editor*, active atunci când se doreşte afişarea (deschiderea) sau modificarea rezultatelor. Operaţia este posibilă fie din *meniul rapid*[7], fie din meniul *Edit*, folosind comanda *SPSS Pivot Table Object, Edit/Open* (vezi figura 1.7) şi respectiv *SPSS Chart Object, Open* (vezi figura 1.8). Atunci când rezultatele nu sunt afişate în tabele pivot, modificarea este posibilă în fereastra *Text Editor Output*.

7. *Meniul rapid* se activează cu butonul din dreapta al mouse-ului.

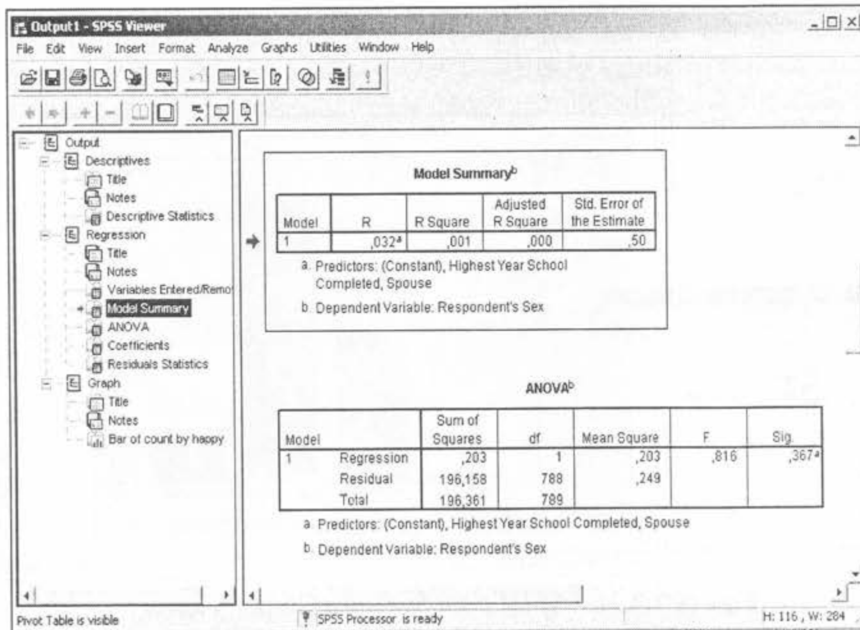


Figura 1.6 *Fereastra* Output Viewer

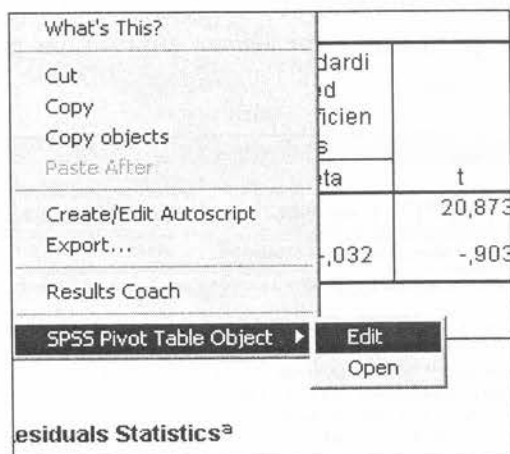


Figura 1.7 *Meniul rapid din fereastra* Output Viewer

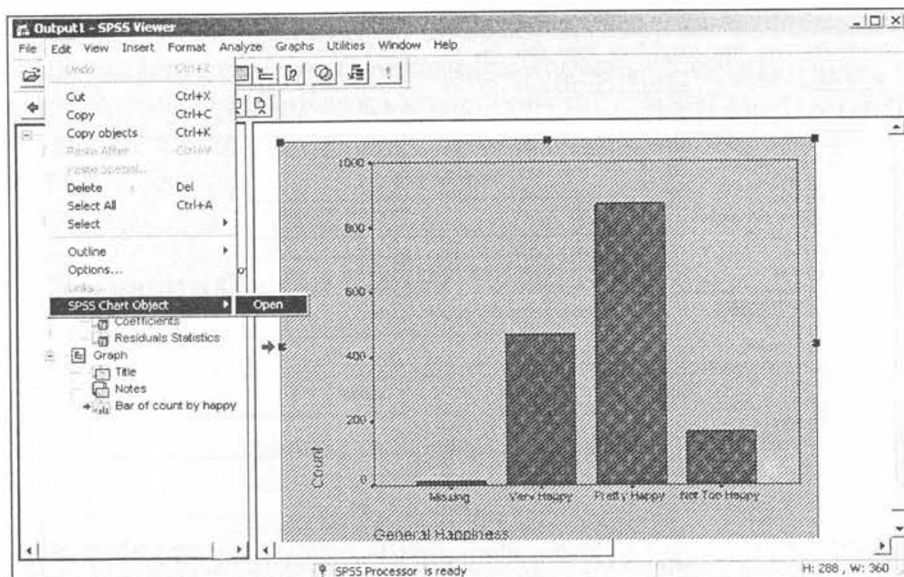


Figura 1.8 *Meniul* Edit *din fereastra* Output Viewer

Din fereastra *Output Viewer*, prin meniul rapid, comanda *Create/Edit Autoscript*, se deschide o nouă fereastră *Scripts Viewer* care este similară unei ferestre Visual Basic şi în care sunt afişate subrutinele programului generat (vezi figura 1.9).

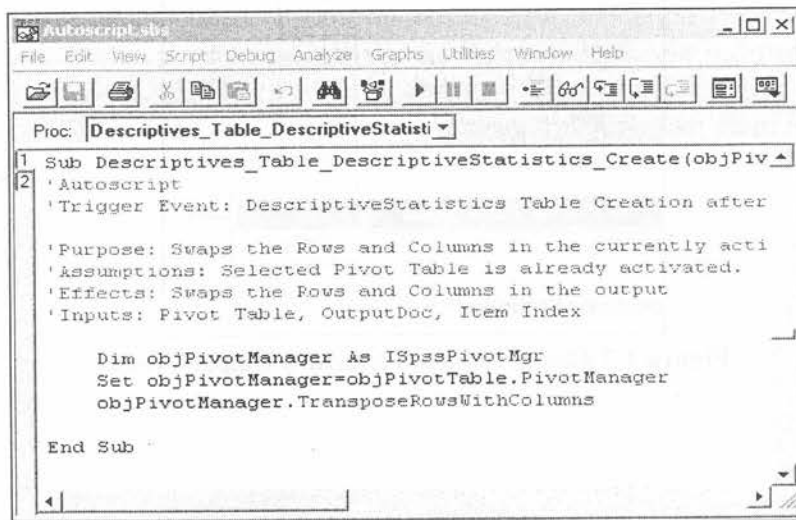


Figura 1.9 *Fereastra* Autoscript

Trebuie făcută precizarea că atât fereastra *Output*, cât şi ferestrele asociate ei plasează în bara de sarcini/task-uri *butoane* distincte, corespunzătoare operaţiilor de obţinere, extragere sau editare a rezultatelor (vezi figura 1.10).



Figura 1.10 *Bara de task-uri*

1.3.4 Obiecte de control în ferestrele SPSS

Interfaţa SPSS oferă o serie de elemente, numite *obiecte de control*, care simplifică/uşurează dialogul utilizatorului cu sistemul de calcul. Aceste elemente se regăsesc în toate produsele program care rulează sub sistemul de operare Windows. În figura 1.11 sunt prezentate principalele obiecte de control din fereastra cadrului de pagină *General*, subordonată comenzii *Options* din meniul *Edit*.

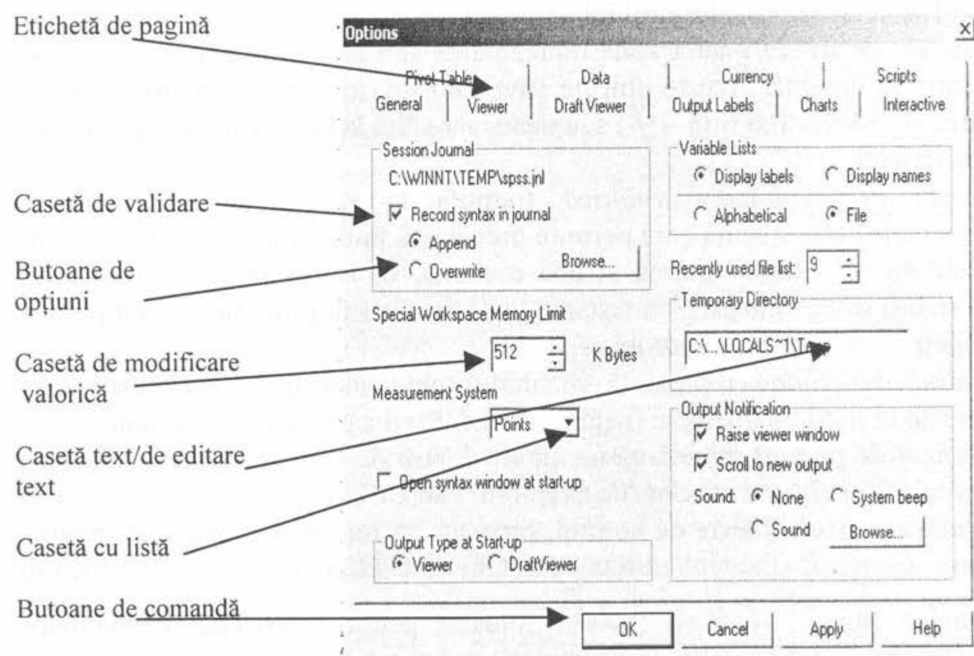


Figura 1.11 *Obiecte de control în ferestrele SPSS*

Caseta de text/de editare este o zonă în care utilizatorul, folosind comenzi de editare, poate prelua, introduce sau modifica text (nume de foldere, fişiere, variabile, expresii etc.).

Caseta cu listă este o zonă în care sistemul afişează total sau parţial un grup de elemente din care se poate selecta unul singur.

Caseta combinată conţine caracteristicile atât ale unei casete text, cât şi ale unei casete cu listă, existând posibilitatea introducerii/modificării de text sau selectarea unui element din lista derulantă ascunsă.

Butoanele de comandă sunt zone sub formă de dreptunghi în care apare numele butonului. Acest nume sugerează funcţia butonului. Unele butoane au în plus trei puncte de suspensie ce indică deschiderea unei ferestre de dialog suplimentare. Fiecare fereastră are un buton implicit, cel care are marginile umbrite (de exemplu, pentru fereastra cadrului de pagină *General*, butonul implicit este *OK*).

Butoanele de opţiuni (numite şi butoane radio) sunt elemente care se exclud reciproc, limitând utilizatorul la selectarea unei singure variante din cele posibile. Se prezintă sub forma unui cerc cu un text explicativ în dreapta lor. Butonul selectat are în interiorul lui un punct.

Casetele de validare sunt zone reprezentate sub formă de pătrat cu un text explicativ la dreapta. Aceste obiecte sunt folosite, în general, pentru a indica valoarea de adevăr (cu bifă – √) sau neadevăr (fără bifă) a condiţiei impuse de funcţia casetei.

Caseta de modificare valorică (numită şi casetă de incrementare/ decrementare) este o zonă care permite precizarea unui număr sau a unei valori. Casetele au o valoare minimă şi una maximă. În acest interval, utilizatorul poate stabili orice valoare prin tastarea unui număr sau prin clic repetat pe una din săgeţi (superioară sau inferioară).

Cadrul de pagină reprezintă rezultatul unui mecanism prin care sunt organizate în module distincte (pagini) elemente cu caracteristici comune.

Eticheta de pagină reprezintă un şir invariabil de caractere (text), afişat pe elementul căruia îi este asociat (de exemplu, *numele cadrului de pagină*).

În afara acestor obiecte de control, prezente în fereastra cadrului de pagină *Options*, *General*, dialogul utilizator-sistem este asigurat şi prin intermediul butoanelor săgeţi: ▸ şi ◂, folosite pentru transferul elementelor selectate dintr-o zonă în alta. În general, aceste săgeţi sunt prezente în ferestrele de dialog (vezi figura 1.12).

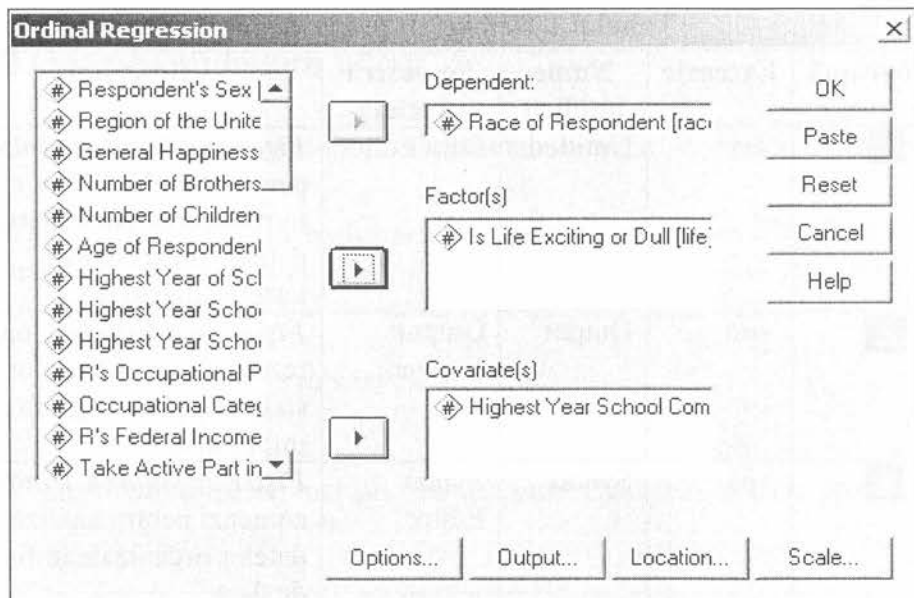


Figura 1.12 *Obiecte de control în ferestrele SPSS*

1.4 Gestiunea fişierelor SPSS

1.4.1 Tipuri de fişiere

SPSS utilizează patru tipuri de fişiere: date, rezultate, grafice şi sintaxă. Tabelul 1.1 prezintă tipurile de fişiere cu care se lucrează într-o sesiune SPSS şi principalele lor caracteristici.

Fişierul de date este specific ferestrei *Data Editor* şi este identificat prin extensia *.sav*[8].Numele implicit este *Untitled*, urmat de un număr care semnifică al câtelea fişier de date este creat în sesiunea curentă de lucru. Pentru o identificare rapidă a fişierelor de date sunt recomandate nume care să sugereze conţinutul informaţional ori apartenenţa la o anumită aplicaţie sau un anumit utilizator. Comenzile statistice şi graficele operează asupra datelor organizate în astfel de fişiere.

8. Pentru versiunile *SPSS* sub MS-DOS, extensia fişierelor de date este *.sys*.

Tabelul 1.1 *Tipuri de fişiere SPSS*

Pictogramă	Extensie	Nume implicit	Fereastră asociată	Descriere
	.sav	Untitled	Data Editor	***Fişier de date***. Este folosit pentru definirea, introducerea sau editarea datelor şi executarea testelor statistice.
	.spo	Output	Output Viewer	***Fişier de rezultate***. Conţine rezultatele prelucrărilor statistice (tabele, grafice şi informaţii)
	.sps	Syntax	Syntax Editor	***Fişier de sintaxă***. Conţine comenzi pentru analiza datelor organizate în fişiere de date
	.cht	Chart	Chart Editor	***Fişier de grafice***. Conţine reprezentarea grafică a datelor din fişierul de date

Fişierul de rezultate este specific ferestrei de rezultate *Output Viewer* şi este recunoscut după extensia *.spo*. Numele implicit este *Output*, urmat de numărul de ordine al fişierului. Pot fi mai multe fişiere de rezultate stocate în aceeaşi fereastră de rezultate. Fişierul activ la un moment dat este specificat în fereastra de rezultate prin simbolul → (săgeată la dreapta) în dreptul său, atât în cadrul din stânga, cât şi în cel din dreapta ferestrei *Output Viewer* (vezi figura 1.6).

Fişierul de grafice este asociat ferestrei de grafice şi are extensia *.cht*. Opţiunea grafică, într-o procedură statistică, permite reprezentarea rezultatelor într-o fereastră specifică, *Chart Editor*.

Fişierul de sintaxă este subordonat ferestrei *Syntax Editor* şi are extensia *.sps*. Un fişier de sintaxă reprezintă un ansamblu de comenzi care realizează analiza informaţiilor stocate într-un fişier de date.

1.4.2 Operaţii cu fişiere SPSS

Crearea fişierelor SPSS presupune exploatarea facilităţilor oferite de ferestrele cărora le sunt subordonate.

Salvarea unui fişier se realizează prin pictograma *Save* din bara cu instrumente *Standard* sau cu ajutorul comenzilor *Save* sau *Save As* din meniul *File*. Aceste comenzi deschid fereastra *Save Data As* (vezi figura 1.13), în care se pot stabili:

- numele fişierului (*File name*);
- tipul fişierului (*Save as type*);
- locaţia în care să aibă loc salvarea (*Save In*): directorul/folderul/calea de directoare/foldere, inclusiv un director creat *pe loc* cu pictograma *Create New Folder* .

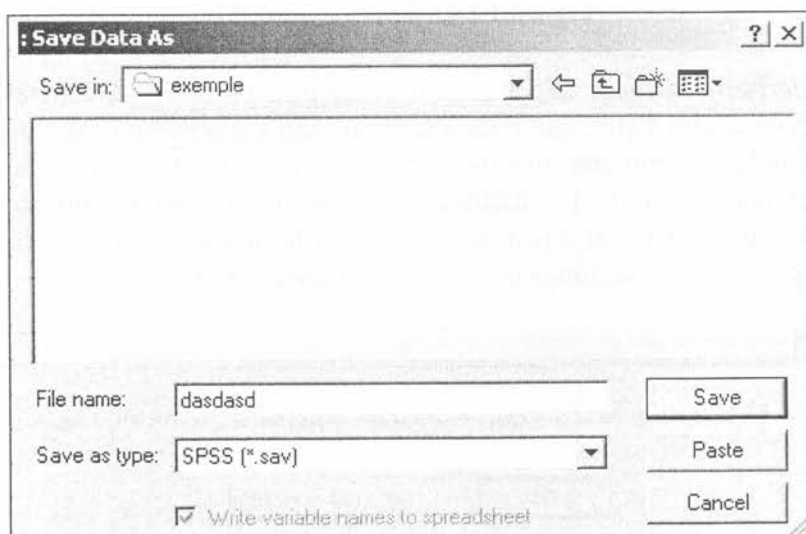


Figura 1.13 *Fereastra* Save Data As

Pentru *deschiderea* fişierelor se foloseşte pictograma *Open* din bara cu instrumente *Standard* sau comanda *Open* din meniul *File*. Aceste opţiuni deschid fereastra *Open File* (vezi figura 1.14), în care este posibilă şi *localizarea* unui fişier (în zona *Look in:)*, dacă acesta nu se află în folderul curent în acel moment.

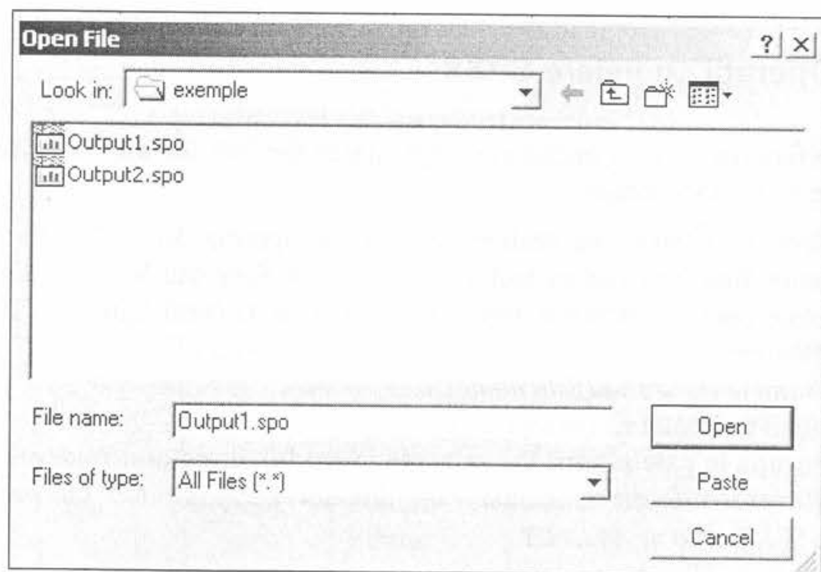


Figura 1.14 *Fereastra* Open File

Închiderea unui fişier SPSS se realizează prin butonul ☒, asociat ferestrei subordonate acelui fişier, sau o dată cu terminarea unei sesiuni de lucru SPSS, prin comanda *Exit* din meniul *File* (caz în care se închid toate fişierele deschise în respectiva sesiune). La închidere, SPSS interoghează utilizatorul dacă salvează sau nu fişierul creat ori modificările realizate într-un fişier creat anterior şi deschis în sesiunea curentă (vezi figura 1.15).

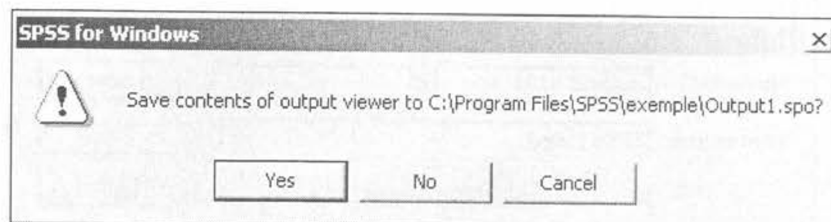


Figura 1.15 *Butoanele subordonate operaţiei de salvare*

1.4.3 Barele cu instrumente SPSS

Barele cu instrumente (*toolbars*) se constituie din scurtături create pentru cele mai apelate comenzi din meniurile *SPSS*.

Principalele bare de instrumente în *SPSS* sunt:

- Data Editor;

- Syntax Editor;
- Viewer Standard;
- Viewer Outlining;
- Draft Viewer Standard;
- Draft Viewer Formatting;
- Chart Standard;
- Chart Formatting;
- Script Editor.

Pe lângă aceste instrumente, utilizatorul poate să-şi definească bare noi, prin activarea butonului *New Toolbar* (vezi figura 1.16).

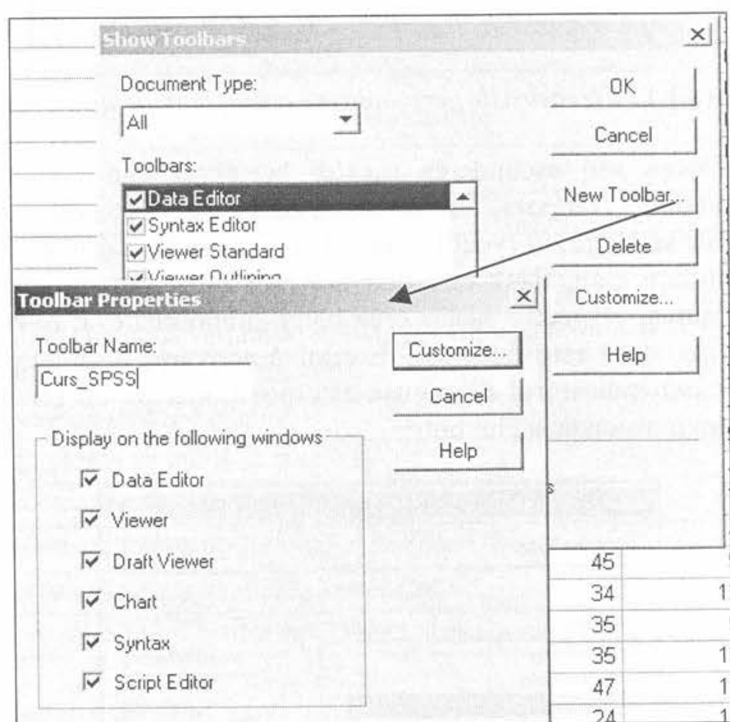


Figura 1.16 *Crearea unei bare de instrumente*

Din fereastra *Toolbar Properties* se stabileşte numele noii bare de instrumente, iar în fereastra *Customize Toolbar* (deschisă cu butonul de comandă *Customize*) se stabilesc butoanele/pictogramele ce se vor afişa în noua bară (vezi figura 1.17).

Aşa cum se poate adăuga o bară de instrumente, tot la fel poate fi ştearsă, dacă se foloseşte butonul de comandă *Delete* din fereastra *Show Toolbars* (vezi figura 1.16).

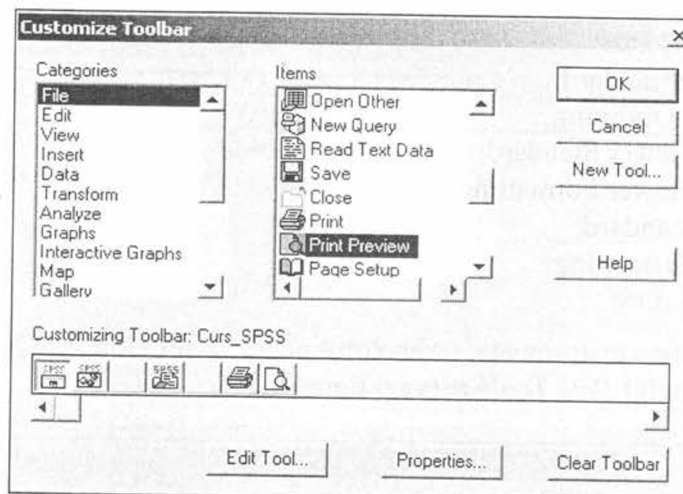


Figura 1.17 *Fereastra de personalizare a unei bare de instrumente*

Pentru afişarea sau ascunderea acestor bare/linii, din meniul *View* se selectează comanda *Toolbars*. Din fereastra *Show Toolbars*, din lista ascunsă *Document Type* se alege *All* (vezi figura 1.18). Fiecare bară este prevăzută cu o casetă de validare care, dacă este activată (are bifa) sau nu (nu are bifa), determină afişarea, respectiv ascunderea barei subordonate. Caseta de dialog *Show Tool Tips*, dacă este selectată, determină activarea *help*-ului contextual care, atunci când indicatorul de mouse este poziţionat pe un buton, afişează numele comenzii asociate acelui buton.

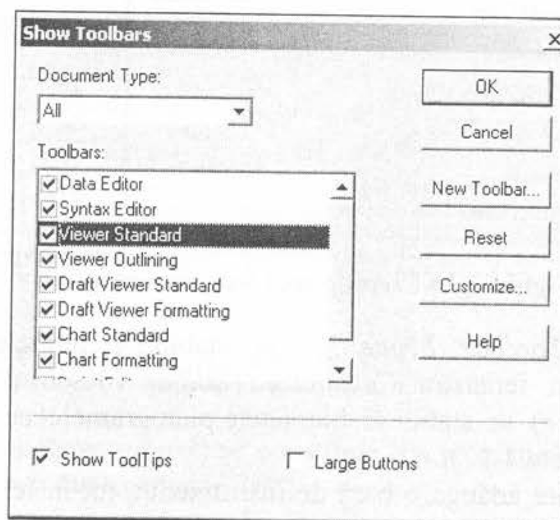


Figura 1.18 *Fereastra* Show Toolbars

Prezentăm în continuare cele mai folosite butoane, din cele mai apelate bare de instrumente – Data Editor şi Chart Standard.

Data Editor Toolbar. Această bară de instrumente apare când fereastra *Data Editor* este activată. Ea conţine butoane-scurtături pentru cel mai frecvent utilizate acţiuni: deschiderea sau salvarea unui fişier, tipărirea datelor şi rezultatelor, introducerea datelor etc. (vezi tabelul 1.2).

Tabelul 1.2 *Butoane în bara de instrumente* Data Editor

Buton	Efect
	Deschiderea unui fişier de date (*.sav*) sau de rezultate (*.spo*)
	Salvarea unui fişier de date (*.sav*) sau de rezultate (*.spo*)
	Tipărirea fişierelor de date sau de rezultate
	Anularea acţiunii precedente
	Revenirea la acţiunea precedentă
	Accesarea casetei de dialog *Chart*
	Caută anumite cazuri (rând)
	Caută o anumită variabilă (coloană) şi afişează informaţii despre aceasta
	Caută date (numai în *Data View*)
	Inserează un caz (un rând)
	Inserează o variabilă (o coloană)
	Accesează caseta de dialog *Split File*
	Accesează caseta de dialog *Weight Cases*
	Accesează caseta de dialog *Select Cases*
	Accesează seturi pentru caseta de dialog *Variables*

Chart Standard Toolbar. Linia *Chart Standard* apare în fereastra editorului de grafice (SPSS Chart Editor) şi conţine butoane-scurtături pentru cel mai adesea utilizate acţiuni (vezi tabelul 1.3).

Tabelul 1.3 *Butoane în bara de instrumente* Chart Standard Toolbar

Buton	Efect
	Fill Pattern – permite stabilirea „modelului" de umplere
	Schimbă culorile pentru elementele din grafic
	Schimbă tipul punctelor

	Accesează bara de dialog *Line Type*
	Accesează caseta de dialog *Bar Type*
	Accesează caseta de dialog *Bar Label Style*
	Accesează caseta de dialog *Point Interpolation*
	Modifică stilul textului din grafic
	Accesează caseta de dialog *3-D Axis Rotation*
	Accesează caseta de dialog *Swap axes*
	Accesează caseta de dialog *Explode Slice*
	Accesează caseta de dialog *Break Lines*
	Modifică opțiuni (*Bar/Line/Area Options*)
	Setează modul *spin* (*3-D Scaperplot*)

1.4.4 Meniurile în SPSS

Bara meniu conține mai multe meniuri pe care, succint, le prezentăm în continuare.

File. Acest meniu este folosit pentru realizarea operațiilor curente asupra fișierelor: deschidere pentru crearea unui nou fișier (*New*), deschiderea unui fișier existent (*Open*), salvare (*Save* sau *Save As*), tipărire (*Print*), vizualizare înainte de tipărire (*Print Preview*) etc. (vezi figura 1.19). Tot din acest meniu se asigură închiderea sesiunii de lucru SPSS (*Exit*).

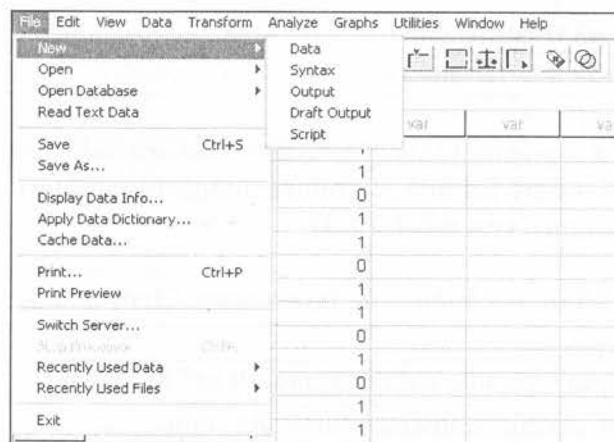


Figura 1.19 *Meniul* File

Edit. Comenzile acestui meniu operează în ferestrele de rezultate şi de sintaxă şi asigură executarea operaţiilor de copiere şi/sau mutare (*Copy*, *Cut*, *Paste*, *Paste Variables*), ştergere (*Delete*) şi căutare rapidă (*Find*) (vezi figura 1.20).

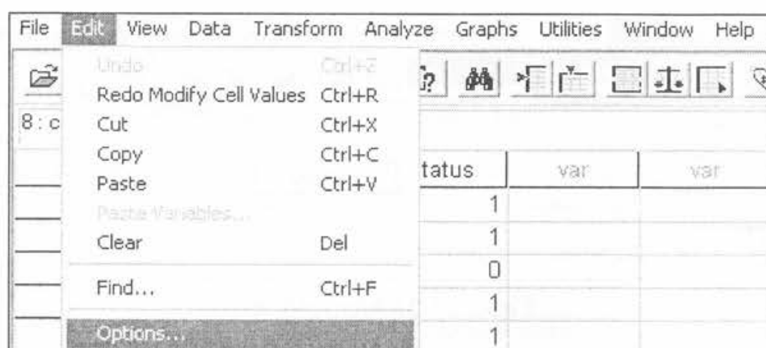


Figura 1.20 *Meniul* Edit

De asemenea, din acest meniu pot fi definite o multitudine de *opţiuni* care personalizează mediul de desfăşurare a sesiunii de lucru (vezi figura 1.21).

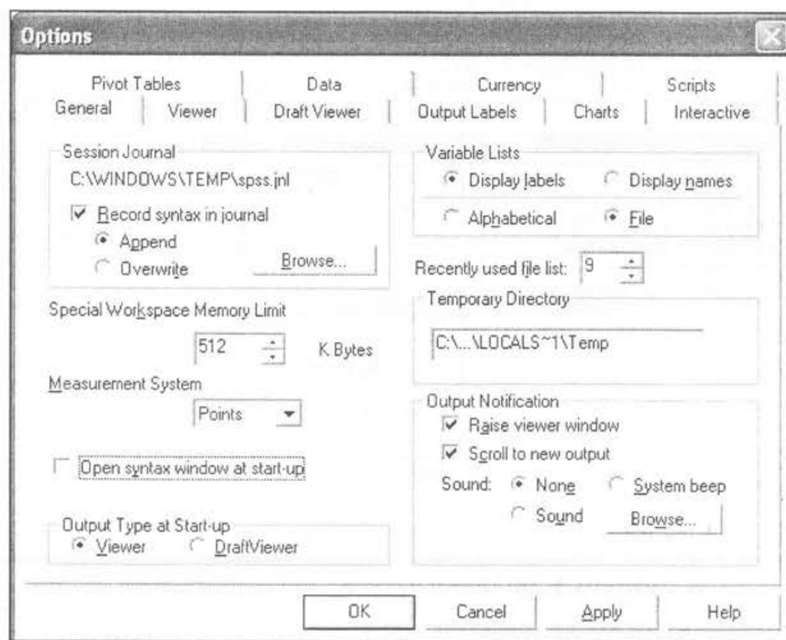


Figura 1.21 *Fereastra* Options *din meniul* Edit

View. Acest meniu, prin comenzile subordonate, permite afişarea sau neafişarea barei de stare (*Status Bar* – plasată în partea de jos a monitorului,

deasupra barei de task-uri şi sub foile *Data View* şi *Variable View*), a altor bare cu instrumente de lucru (*Toolbars*) sau a grilelor/liniilor din foile ferestrei de editare (*Grid Lines*) (vezi figura 1.22).

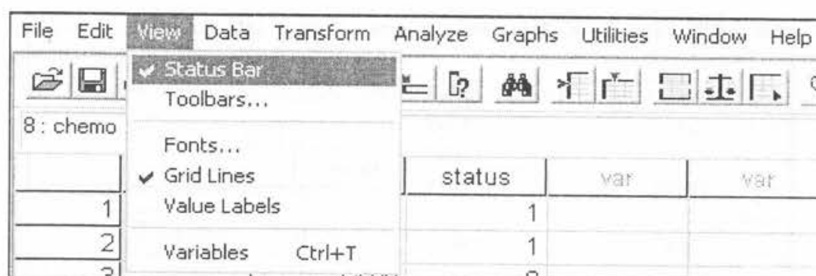


Figura 1.22 *Meniul* View

Comanda *Fonts* deschide fereastra *Font* în care se pot modifica fişierele ce conţin fonturi (*Font*: Arial, Helvetica, Letter Gothic, Tahoma, Times New Roman etc.), stilul fonturilor (*Font Style*: Italic, Bold etc.), precum şi mărimea fonturilor (*Size*: 6, 10, 11, 14 etc.) (vezi figura 1.23).

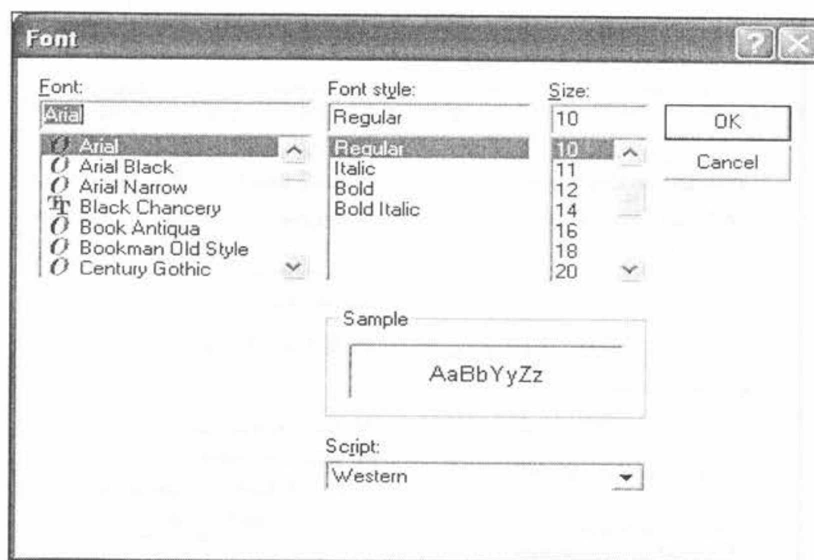


Figura 1.23 *Fereastra* Font *din meniul* View

Data. Prin comenzile acestui meniu este „afectat" conţinutul ferestrei *Data Editor*. Poate fi stabilit formatul de afişare a datelor calendaristice şi a timpului (*Define Dates*: zi, zile lucrătoare, săptămână, lună, oră, minut, secundă etc.), se pot introduce variabile şi cazuri (*Insert Variable*, *Insert Case*), pot fi localizate

rapid, conform unei numerotări, cazurile (*Go to Case*). Cazurile pot fi sortate (*Sort Case*) crescător (*Ascending*) sau descrescător (*Descending*). Facilităţi deosebite sunt cele care permit fuzionarea fişierelor (*Merge Files*) sau splitarea/împărţirea lor (*Split File*). De asemenea, este posibilă selectarea cazurilor (*Select Cases*) şi stabilirea ponderii cazurilor (*Weight Cases*) (vezi figura 1.24).

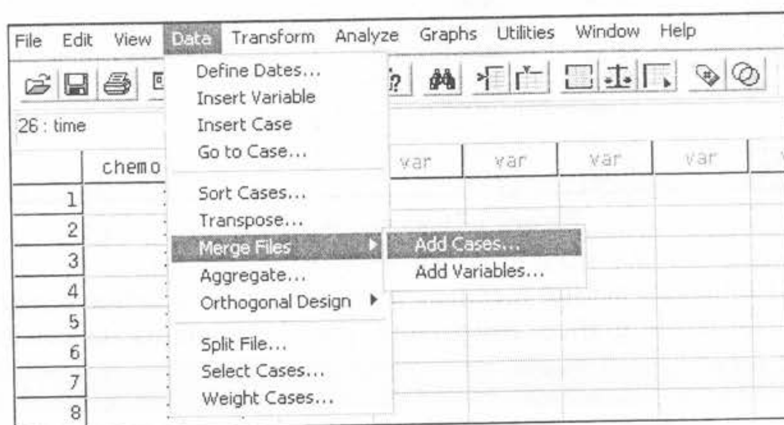


Figura 1.24 *Meniul* Data

Transform. Acest meniu este utilizat pentru transformarea datelor sau pentru crearea unor variabile noi (vezi figura 1.25).

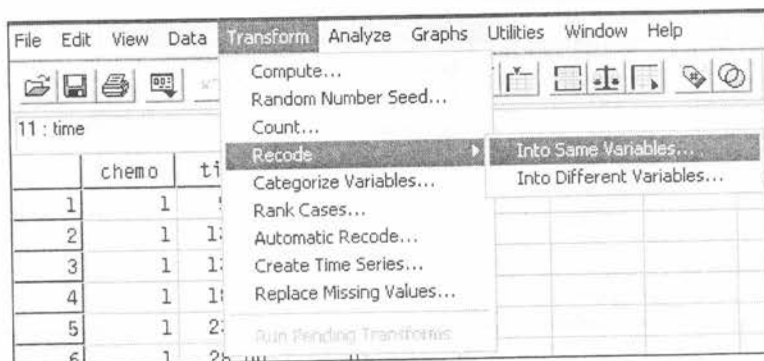


Figura 1.25 *Meniul* Transform

Analyze. Acest meniu este folosit pentru realizarea procedurilor statistice (vezi figura 1.26).

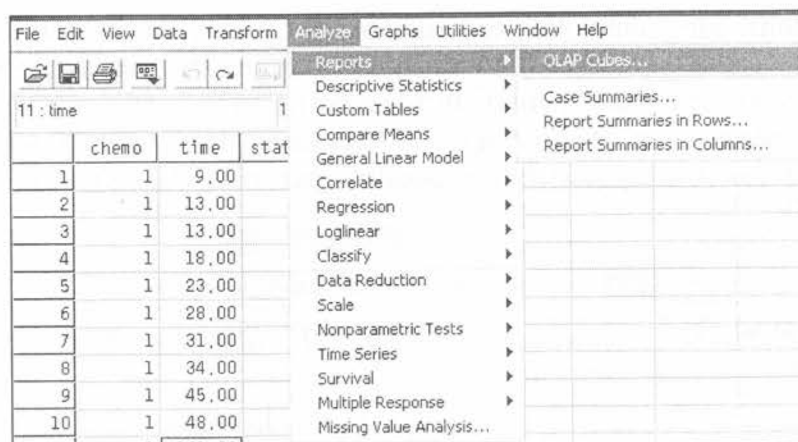


Figura 1.26 *Meniul* Analyze

Graphs. Comenzile acestui meniu sunt folosite pentru a obține reprezentarea datelor sub formă de grafice: histograme, puncte, diagramă de structură etc. (vezi figura 1.27)

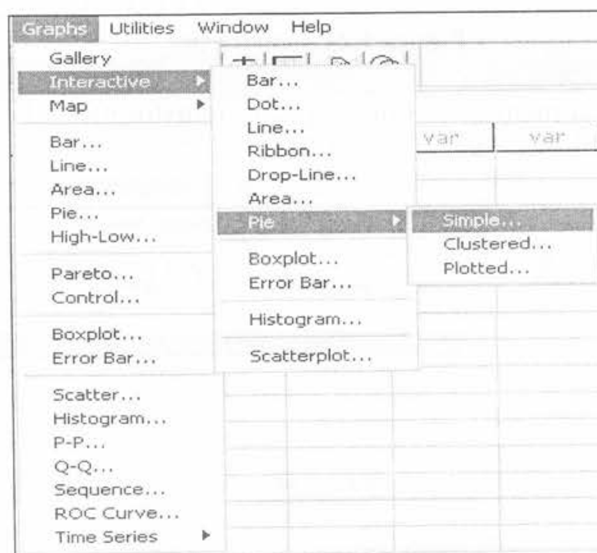


Figura 1.27 *Meniul* Graphs

Utilities. Acest meniu reunește, sub forma unui index al comenzilor, cele mai utilizate instrumente, cu o scurtă descriere a acestora: informații privind variabilele curente (*Variables*), informații despre fișierele disponibile (*File Info*), definirea și utilizarea seturilor (*Define Sets*, *Use Sets*) (vezi figura 1.28).

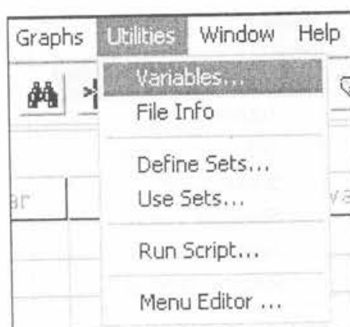


Figura 1.28 *Meniul* Utilities

De asemenea, din acest meniu este posibilă lansarea scripturilor (*Run Script*) şi activarea meniului de editare pentru configurarea personalizată a meniurilor (*Menu Editor*) pentru ferestrele *Data Editor*, *Viewer*, *Script* şi *Syntax* (vezi figura 1.29).

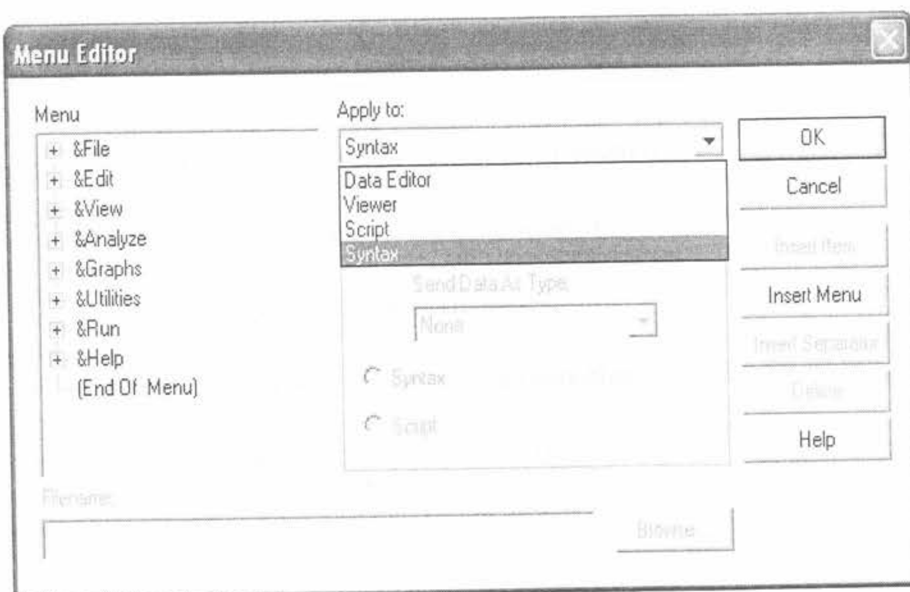


Figura 1.29 *Fereastra* Menu Editor

Windows. Comenzile meniului Windows asigură comutarea între ferestrele diferitelor fişiere deschise, precum şi controlul aranjării acestora pe ecran.

Help. Acest meniu permite familiarizarea cu SPSS. Cele mai utilizate opţiuni sunt *Topics*, care afişează un meniu contextual în funcţie de subiectul precizat de utilizator, şi *Tutorial*, care oferă asistenţă în învăţarea SPSS (vezi figura 1.30).

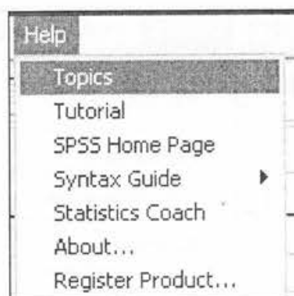


Figura 1.30 *Meniul* Help

Caracteristicile SPSS descrise mai sus ne prezintă un produs program care oferă facilități de lucru performante pentru o gamă largă de utilizatori care folosesc statistica, fie în activitatea practică, fie în cercetarea științifică.

CAPITOLUL 2

ELEMENTE CONCEPTUALE ŞI METODOLOGICE DE STATISTICĂ

- **Obiectul de studiu, metoda şi scopul statisticii**
- **Ipostaze ale statisticii – ştiinţă şi metodă**
- **Particularităţi ale metodei statisticii**
- **Etape ale procesului cunoaşterii statistice**
- **Noţiuni fundamentale ale statisticii**
- **Notaţii**

2.1 Obiectul de studiu, metoda şi scopul statisticii

Statistica studiază fenomene de masă, de tip stochastic. Astfel de fenomene îşi au originea în existenţa colectivităţilor, ansambluri formate dintr-un număr mare de elemente unite între ele prin trăsături esenţial comune, care se manifestă şi pot fi cunoscute numai la nivelul întregului. Trăsăturile esenţial comune exprimă valoarea medie, adică normală, predominantă, reflectată de majoritatea elementelor unei colectivităţi.

De exemplu, populaţia de o anumită vârstă este distribuită în funcţie de „înălţime" după modelul cunoscut sub denumirea de „curba normală" sau „clopotul Gauss". Acest model ne arată că majoritatea indivizilor de aceeaşi vârstă au aproximativ aceeaşi înălţime, de la care se abat, progresiv, în plus şi minus, din ce în ce mai puţini indivizi. Pentru majoritatea indivizilor dintr-o populaţie, înălţimea este consecinţa condiţiilor normale în care se dezvoltă acea populaţie. Abaterile de la normal sunt atribuite unui ansamblu de factori care acţionează într-un sens sau altul şi determină perturbări.

Modelul distribuţiei indivizilor după înălţime rămâne acelaşi, indiferent de localitatea sau momentul de observare a populaţiei. Ceea ce diferă de la o populaţie la alta, în alte condiţii de spaţiu sau de timp, sunt parametrii modelului. Fiind rezultanta acţiunii factorilor esenţiali asupra fenomenelor de masă, modelul scoate în evidenţă ceea ce este comun în majoritatea cazurilor.

Totodată, trebuie reţinut că această influenţă este însoţită şi de acţiunea unor factori aleatorii, care duc la apariţia unor perturbări de la ceea ce este normal. Aceste perturbări nu rămân constante în timp şi spaţiu. Ca urmare, am putea şti care este modul de distribuţie a populaţiei după variabila înălţime, indiferent de timpul şi spaţiul de referinţă, dar nu ce înălţime ar înregistra un anume individ din colectivitate.

Prin studiul statistic al fenomenelor se desprind trăsăturile comune, comportarea normală a fenomenelor la nivelul ansamblului, nu al fiecărui individ în parte. Statistica observă fiecare element al unei colectivităţi în variabilitatea sa şi ajunge, prin prelucrarea datelor obţinute din observarea statistică şi compararea rezultatelor prelucrării, la cunoaşterea întregului.

Înregistrând aleatoriu fiecare element al unei colectivităţi, în formele sale particulare de manifestare în timp, în spaţiu şi din punct de vedere calitativ, statistica ajunge, prin metoda sa particulară, să reţină la nivelul întregului numai ceea ce este normal, esenţial pentru toate unităţile colectivităţii observate. Cum ar spune C. Noica, statistica *lasă întregurile să iasă din elementele infinitezimale*[1].

1. Constantin Noica, *Jurnal filozofic*, Ed. Humanitas, Bucureşti, 1990, p. 57.

Statistica este o știință cu un obiect de studiu propriu, o metodă particulară și un scop bine precizat.

Obiectul de studiu al statisticii îl constituie *variația curentă-continuă, în timp, în spațiu și din punct de vedere calitativ a fenomenelor de tip stochastic* din orice domeniu al vieții economico-sociale sau naturale. Particularitatea obiectului de studiu al statisticii este dată de un mod specific de a privi elementele vieții materiale, și anume în mișcarea lor curentă-continuă, în timp, în spațiu și din punct de vedere calitativ.

Fenomenele, cum precizează profesorul Alexandru Bărbat în *Teoria statisticii sociale*, nu pot deveni obiect de studiu al statisticii în forma lor de substanță materială, ci numai sub formă de mișcare, și anume sub formă de mișcări curente-continue privind creșterea, descreșterea, diversificarea și modificările „structurale” ale fenomenelor și proceselor de tip colectiv[2].

Metoda statisticii este definită ca *un ansamblu de principii metodologice, procedee și tehnici care permit producerea informației statistice*, pe baza observării statistice, a prelucrării și analizei datelor statistice, *precum și fundamentarea deciziilor privind starea și variabilitatea colectivităților statistice, în timp, în spațiu și din punct de vedere calitativ.* (Particularitățile metodei statistice sunt prezentate în paragraful 2.3).

Scopul statisticii este cunoașterea fenomenelor de masă, caracterizate prin variabilitate și produse sub semnul incertitudinii. *Vizează*, pe de o parte, *elaborarea informației statistice* necesare fundamentării deciziilor asupra colectivităților statistice, iar pe de altă parte, *descoperirea legilor de variabilitate a fenomenelor ce se produc și evoluează sub semnul incertitudinii.*

2.2 Ipostaze ale statisticii – știință și metodă

În procesul cunoașterii, statistica își manifestă *caracterul său dual*, fiind definită atât ca *disciplină de sine stătătoare*, cât și ca *metodă de cercetare* folosită de către alte discipline științifice pentru descoperirea legilor proprii domeniului lor de studiu.

2. Alexandru Bărbat, *Teoria statisticii sociale*, E.D.P., București, 1972, p. 30.

2.2.1 Statistica – ştiinţă

Ca disciplină ştiinţifică, statistica se subdivide, după scopul cunoaşterii, în statistică descriptivă, statistică inferenţială şi analiză statistică.

Statistica descriptivă vizează *descrierea stării şi variabilităţii unei colectivităţi statistice* după una sau mai multe caracteristici. Realizarea acestui obiectiv presupune culegerea datelor statistice, prelucrarea şi prezentarea lor sintetică, fie sub formă numerică prin indicatori statistici, fie sub formă grafică prin diagrame şi tabele statistice. În raport cu numărul caracteristicilor considerate în planul cunoaşterii, se poate vorbi despre o *statistică descriptivă unidimensională* (statistică descriptivă cu o variabilă), respectiv despre o *statistică descriptivă bidimensională sau multidimensională* (cu două sau mai multe variabile).

Statistica inferenţială a apărut abia după descoperirea legilor probabilistice şi construirea teoriei probabilităţilor ca ştiinţă. *Statistica inferenţială* vizează *estimarea caracteristicilor unei colectivităţi pornind de la cunoaşterea unei colectivităţi parţiale* şi presupune măsurarea incertitudinii rezultatelor şi calcularea riscurilor pe care le implică luarea unei decizii fundamentate pe baza unei informaţii ce nu poate fi exhaustivă. Principalele probleme ale inferenţei statistice sunt *estimarea parametrilor distribuţiei unei colectivităţi* şi *testarea ipotezelor statistice.*

Analiza statistică urmăreşte *descoperirea a ceea ce este permanent, esenţial, legic în variaţia proceselor stochastice şi măsurarea influenţei factorilor care le determină variaţia în timp, în spaţiu şi din punct de vedere calitativ.* În acest scop se folosesc, în principal, analiza de regresie, analiza de corelaţie, ANOVA, analiza seriilor de timp.

2.2.2 Statistica – metodă

Ca metodă, statistica a câştigat în timp un loc important printre metodele ştiinţelor fundamentale: fizică, chimie, biologie, astronomie etc. Astăzi, toate disciplinele ştiinţifice care investighează fenomene de masă (ştiinţe economice, sociologice, agronomice, meteorologia, genetica, medicina etc.) apelează la metoda statistică pentru descoperirea legilor proprii domeniului lor de studiu, a permanenţelor şi tendinţelor care se pot constitui ca elemente de previziune.

De altfel, în condiţiile în care ritmul dezvoltării şi evoluţiei societăţii moderne a imprimat un caracter de masă fenomenelor din domeniul tehnic,

economic, social, al conducerii afacerilor etc., metoda statistică a devenit un instrument indispensabil de cunoaştere.

Caracterul de masă şi variabilitatea unor astfel de fenomene, sub acţiunea factorilor conjuncturali, nu permit cercetarea lor pe baza experienţei, în condiţii deterministe, care să asigure o cunoaştere precisă a lor. De exemplu, fluctuaţia forţei de muncă, fenomen variabil în timp şi depinzând de factori conjuncturali, nu se poate măsura prin experienţă, deoarece nu se pot crea condiţii de producţie, de muncă etc. identic reproductibile în timp şi spaţiu. În astfel de situaţii, când fenomenele de masă se produc în condiţii de incertitudine, se poate folosi metoda statistică.

Metoda statistică ne conduce la *concluzii probabile, nu absolut sigure*. Nu permite, după aprecierea lui A. Piatier[3], *să se afirme certitudinea, ci să se cerceteze limitele de incertitudine*. Prin statistică se calculează aceste limite şi se elaborează probabilitatea de reapariţie a evenimentelor considerate. Folosirea statisticii, în calitate de metodă de cercetare, de către alte ştiinţe se fundamentează tocmai pe această posibilitate oferită de metoda sa particulară, de a descoperi legile de manifestare a fenomenelor de masă, care se desfăşoară în condiţii de incertitudine, pe baza frecvenţei şi regularităţii cu care aceste evenimente au apărut în trecut.

Este edificator, în acest sens, exemplul considerat de Milton Smith[4] cu privire la probabilitatea de producere a două evenimente: apariţia soarelui în ziua de mâine şi apariţia unei zile de întâi ianuarie, în viitor, mai călduroasă decât o zi de întâi aprilie. Observând frecvenţa de apariţie a celor două fenomene, se poate constata că probabilitatea de apariţie în viitor a unui fenomen este în directă legătură cu gradul de constanţă sau de inconstanţă cu care s-a manifestat în trecut.

În cazul dat, cu privire la apariţia soarelui se constată o permanenţă, deci putem fi siguri că şi mâine, atât timp cât există sistemul nostru solar, soarele va apărea; cu privire la apariţia unei zile de întâi ianuarie mai călduroasă decât o zi de întâi aprilie nu mai putem fi tot atât de siguri. Producerea evenimentului în viitor este incertă. Statistica măsoară gradul de incertitudine în producerea unui eveniment în raport cu frecvenţa de apariţie în trecut a evenimentului şi calculează limitele acesteia pentru un anumit risc de a nu se păstra, în viitor, aceleaşi condiţii.

3. A. Piatier, *Statistique descriptive et initiation a l'analyse*, Themis, Paris, 1962, p. 5.
4. Milton Smith, *Ghid simplificat de statistică pentru psihologie și pedagogie*, E.D.P., Bucureşti, 1971, pp. 15-16.

2.2.3 Diversificarea statisticii

Conturată inițial ca știință a „treburilor statului", statistica și-a lărgit treptat sfera de observație, de la domeniul demografic la domeniul economic, social, fizic etc., dezvoltându-se ca ansamblu de statistici aplicate.

Diversitatea domeniilor de existență a fenomenelor de masă a făcut necesară diversificarea statisticii în raport cu natura fenomenelor și proceselor observate. S-au constituit și dezvoltat *statistici specializate* care observă și studiază mișcările curente-continue ale vieții concrete din domeniile respective. Astfel, sunt cunoscute: *Statistica demografică*, *Statistica economică*, *Statistica juridică*, *Statistica medicală*, *Fizica statistică*, *Statistica matematică*, *Statistica taxonomică* etc.

Procesul de diversificare a statisticii continuă; de dată relativ recentă poate fi considerată *Statistica informațională*, la care o contribuție deosebită și-au adus Onicescu și școala sa.

Totodată, asistăm la apariția unor discipline dezvoltate din aplicarea statisticii, alături de alte metode, în investigarea unor domenii de studiu, cum ar fi cazul *econometriei,* rezultată din aplicarea statisticii și matematicii în investigarea fenomenelor din economie.

2.3 Particularități ale metodei statisticii

Epistemologia stabilește că orice disciplină științifică își are metoda sa generală și particulară, și la rândul ei fiecare metodă este legată în mod specific de un anumit obiect al cercetării.

Particularitatea procesului cunoașterii statistice este dată de obiectul său de studiu și, implicit, de metoda sa particulară care s-a dezvoltat în funcție de specificitatea obiectului de studiu.

Metoda particulară a statisticii se bazează pe un *raționament deductiv--inductiv iterativ* și este definită de un *ansamblu de principii metodologice, procedee și tehnici de lucru* folosite în investigarea fenomenelor observate.

2.3.1 Particularități ale raționamentului statistic

În procesul cunoașterii statistice se utilizează cele două tipuri de raționament ale metodei științifice: deductiv și inductiv.

Metoda deductivă procedează de la general la particular şi utilizează în special raţionamentul matematic: se stabilesc ipotezele generale asupra unei probleme şi se deduc, prin raţionament logic, anumite proprietăţi particulare.

Metoda inductivă presupune procesul invers: se pleacă de la observaţii particulare asupra unor fenomene şi se ajunge la formularea unor reguli generale.

Procesul cunoaşterii statistice începe cu emiterea ipotezelor care implică proprietăţi observabile, care se verifică prin analiza datelor înregistrate, apoi, folosindu-se un ciclu deductiv-inductiv succesiv, se generalizează mai departe ipotezele verificate, *procesul cunoaşterii statistice fiind un proces iterativ.*

O altă particularitate a procesului cunoaşterii statistice a fenomenelor şi proceselor constă în tratarea acestora ca un întreg structural. Statistica pleacă de la individual la întreg, fiecare element al colectivităţii este observat, fără a fi izolat de întreg, cunoaşterea întregului rezultând din structurarea elementelor componente după o ordine a variaţiei lor în timp, în spaţiu şi din punct de vedere calitativ. La nivelul întregului se reţine numai ceea ce este generat de cauze comune, adică numai ceea ce este normal (purtat de majoritatea elementelor), esenţial, permanent în variaţia curentă-continuă a fenomenelor observate; abaterile de la ceea ce este normal, datorate influenţei unor factori neesenţiali asupra unui element oarecare, se compensează la nivelul întregului.

2.3.2 Principii metodologice ale statisticii

Principiile metodologice care particularizează metoda statisticii sunt: observarea faptică şi exprimarea numerică.

Observarea faptică. Prin natura lor, elementele unei colectivităţi economico-sociale nu pot fi observate, măsurate şi înregistrate în condiţii de laborator, prin experienţă. Specifică acestora este observarea faptică, proces complex ce presupune obţinerea datelor privind colectivităţile economico-sociale. Acest proces implică *observarea, măsurarea şi înregistrarea fiecărui element component al colectivităţii sub aspectul caracteristicilor cuprinse într-un program de observare.* Principiul observării faptice cere observarea elementelor acolo unde ele există şi sub forma în care acestea există în timpul producerii lor.

Exprimarea numerică. Măsurarea fenomenelor şi proceselor observate de statistică, datorită caracterului lor de masă, necesită exprimarea numerică. Un fenomen compus dintr-un număr mare de cazuri nu poate fi cunoscut numai sub formă atributivă. De exemplu, nu este suficient să spunem despre efectivul angajaţilor unei firme că este mare sau mic, trebuie să precizăm numeric câţi angajaţi are firma.

Expresia cantitativă, numerică a fenomenelor şi proceselor permite prelucrarea, analiza şi sinteza datelor obţinute prin observarea faptică a fenomenelor de masă. Prelucrarea datelor, pentru astfel de fenomene, este posibilă numai sub formă numerică, fără să deducem de aici că statistica ar studia numai latura cantitativă. Înregistrând şi prelucrând date numerice, statistica poate să constate, prin compararea rezultatelor obţinute, modificările de esenţă calitativă ce se produc în mişcarea fenomenelor. Folosirea expresiei numerice face posibil calculul parametrilor unei distribuţii (de exemplu, valoarea medie, varianţa), al coeficienţilor de corelaţie etc., facilitează comparările şi elaborarea modelelor de evoluţie a fenomenelor.

2.4 Etape ale procesului cunoaşterii statistice

Procesul cunoaşterii statistice parcurge următoarele etape: punerea problemei, observarea statistică, prelucrarea şi analiza datelor statistice, decizia statistică. În fiecare etapă se aplică procedee şi tehnici specifice, înlesnind observarea şi prelucrarea datelor statistice, precum şi testarea şi analiza informaţiilor statistice.

Etapele procesului cunoaşterii statistice sunt prezentate în figura 2.1.

Punerea problemei presupune definirea problemei în termeni precişi, indicându-se scopul şi aria de investigaţie (fenomenul sau procesul de observat), precum şi variabilele ce se cer studiate.

În această etapă, se efectuează documentarea teoretică şi faptică asupra fenomenului de observat, se emit ipotezele de lucru, se aleg metodele de investigare, se elaborează planul de cercetare.

Observarea statistică este etapa în care se înregistrează caracteristicile elementelor unei colectivităţi, se obţine materialul faptic. De calitatea acestuia depinde esenţial autenticitatea informaţiei statistice.

Culegerea datelor statistice se poate realiza fie prin procedee de înregistrare totală, adică înregistrarea exhaustivă a unei populaţii folosind, de exemplu, recensământul, fie prin procedee de înregistrare parţială, adică înregistrări la nivelul unui eşantion, folosind, de exemplu, anchete prin sondaj.

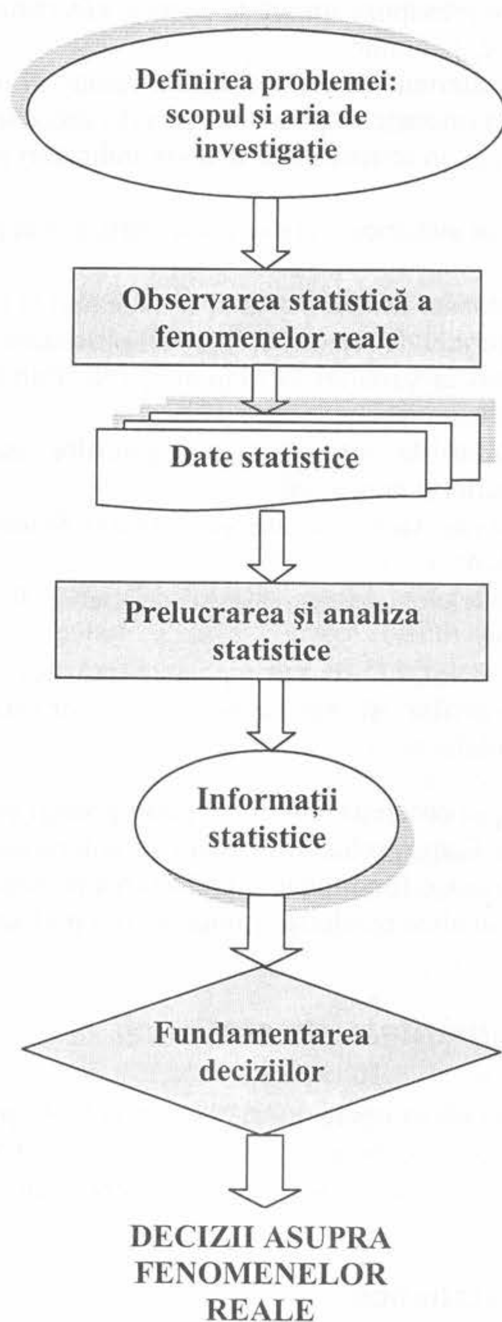


Figura 2.1 *Etapele procesului cunoaşterii statistice*

Prelucrarea statistică presupune un set de operaţii efectuate prin procedee şi tehnici de lucru specifice, şi anume:

- sistematizarea materialului faptic brut obţinut în etapa observării statistice. Această operaţie se poate realiza prin procedee de centralizare şi grupare statistice, în urma cărora se obţin indicatori primari şi serii de date statistice;
- prezentarea datelor statistice, care se poate realiza prin procedee tabelare şi grafice;
- calcularea indicatorilor derivaţi, cum ar fi indicatori ai tendinţei centrale, ai dispersiei, ai formei de repartiţie, folosind procedeul mediei, varianţei etc., sau indicatori ai variaţiei în timp şi spaţiu, folosind, de exemplu, procedeul indicilor statistici;
- măsurarea gradului de intensitate a legăturilor statistice, folosind procedeul covariaţiei şi corelaţiei;
- măsurarea influenţei factorilor asupra variaţiei fenomenelor, folosind procedeul ANOVA;
- aproximarea modelelor de regresie şi de trend, folosind procedeul ajustării statistice;
- prognoza fenomenelor, folosind extrapolarea statistică;
- estimarea parametrilor şi verificarea ipotezelor statistice, folosind procedee inferenţiale.

Rezultatul prelucrării se concretizează în indicatori primari şi derivaţi, purtători ai informaţiei statistice. Etapa prelucrării datelor se îmbină cu analiza acestora. Procesul cunoaşterii statistice fiind iterativ, prelucrarea pe următoarea treaptă se efectuează numai după analiza rezultatelor obţinute din prelucrarea precedentă.

2.5 Noţiuni fundamentale ale statisticii

Procesul cunoaşterii statistice operează cu o terminologie precisă. Noţiunile, conceptele care formează vocabularul de bază al statisticii sunt: colectivităţi statistice, unităţi statistice, variabile statistice, indicatori statistici.

2.5.1 Colectivităţi statistice

Statistica studiază fenomene de masă, ansambluri finite de elemente care au, esenţial, aceeaşi natură, aceleaşi condiţii şi aceleaşi legi de dezvoltare, adică

sunt statistic omogene. Astfel, de exemplu, pot fi considerate următoarele ansambluri:

- populaţia unei ţări la în momentul unui recensământ;
- produsele fabricate de o întreprindere pe parcursul unui an;
- opiniile electorale înregistrate într-o anchetă.

Astfel de ansambluri sunt cunoscute sub denumirea de *populaţii, mulţimi, colectivităţi*.

Denumirea de *populaţie* pentru colectivităţi statistice derivă din faptul că primele aplicaţii ale statisticii se refereau la domeniul demografic, de unde s-a păstrat şi terminologia de bază. Noţiunea de populaţie statistică, respectiv colectivitate statistică, depăşeşte cadrul strict al demografiei şi se poate referi la cele mai diverse domenii, de la mulţimea indivizilor dintr-o ţară la mulţimea stelelor dintr-o galaxie, la un moment dat, sau la cifra de afaceri obţinută de o firmă într-o perioadă de timp.

O colectivitate statistică, respectiv o populaţie sau un univers statistic, reprezintă o asociaţie de elemente unite între ele printr-o trăsătură esenţial comună, numită omogenitate. Elementele unei colectivităţi statistice pot fi fiinţe, lucruri, precum şi fapte, evenimente referitoare la acestea.

Colectivităţile statistice definesc *„populaţii" reale* care sunt întotdeauna *finite*, în contrast cu *„populaţiile" teoretice, infinite*, studiate de matematică. Teoretic, o colectivitate statistică ar putea fi considerată infinită dacă s-ar admite mulţimea tuturor elementelor care pot să existe sau să se producă în aceleaşi condiţii, cum ar fi, de exemplu, mulţimea produselor care ar putea fi fabricate în aceleaşi condiţii. Practic, întotdeauna, *o colectivitate statistică trebuie să fie bine delimitată*, adică trebuie *să i se precizeze frontierele de delimitare* pentru a face posibilă observarea ei şi pentru a nu crea confuzii în interpretare.

De exemplu, o anchetă cu privire la fertilitatea populaţiei unei ţări sau a unei regiuni ar face să apară rezultate nereale dacă în anchetă nu s-ar preciza populaţia observată, în cazul dat populaţia feminină, deoarece s-ar putea înţelege că rezultatele se referă la întreaga populaţie a zonei, adică s-ar include şi bărbaţii şi copiii.

Omogenitate statistică. Delimitarea colectivităţii statistice se face ţinând seama de omogenitatea statistică a elementelor. *Omogenitatea statistică* a elementelor unei colectivităţi presupune *proprietatea acestora de a fi, esenţial, de aceeaşi natură calitativă, de a aparţine aceluiaşi teritoriu şi aceluiaşi timp* (fie unui moment, fie unui interval de timp). Orice colectivitate statistică, într-o

definire completă, trebuie deci să aibă precizată *omogenitatea sub cele trei aspecte: calitativ, spațiu, timp.*

De exemplu, colectivitatea *Populația României la 18 martie 2002* cuprinde *persoanele* înregistrate pe teritoriul *României* la *momentul de referință* (ora 0 din ziua de 18 martie 2002).

Omogenitatea statistică a unităților unei colectivități nu presupune identitatea acestora. Elementele colectivității sunt esențial de același gen din punctul de vedere al caracteristicilor de definire a colectivității, dar se diferențiază între ele după alte caracteristici, pe care, de asemenea, le poartă toate unitățile colectivității, însă cu valori și intensități diferite. Astfel, la nivelul unei colectivități se poate întâlni o diversitate de manifestare a elementelor, din punctul de vedere al gradului de omogenitate în raport cu anumite variabile de distribuție considerate, având ca rezultat conturarea unor subcolectivități și tipuri.

De exemplu, populația unei țări la un moment dat cuprinde totalitatea indivizilor care trăiesc, în acel moment, pe teritoriul respectiv, dar aceștia se diferențiază între ei după diverse caracteristici pe care le posedă, cum ar fi sexul, vârsta, ocupația etc.

Subcolectivități. Subcolectivitățile sunt grupuri de elemente diferențiate între ele, în cadrul colectivității statistice, *din punctul de vedere al unei caracteristici calitative.* Au un grad de omogenitate mai ridicat față de alte grupuri de elemente din aceeași colectivitate.

De exemplu, în cadrul populației umane, se diferențiază, după sex, două subcolectivități: populația masculină și populația feminină.

Tipuri. Tipurile sunt grupuri omogene de elemente, în cadrul unei colectivități sau subcolectivități, *diferențiate între ele după gradul de intensitate sau de dezvoltare atins de o caracteristică dată.*

De exemplu, într-o colectivitate umană, se disting după vârstă următoarele tipuri: *tânăr* (0-19 ani), *adult* (20-59 ani), *în vârstă* (60 de ani și peste)[5].

Clasificarea colectivităților statistice. Pentru a delimita corect o colectivitate statistică este necesar să se cunoască *natura* și *numărul elementelor* componente, precum și *formele de manifestare* ale acestora.

---

5. În literatura de specialitate sunt specificate și alte limite de vârstă pentru cele trei tipuri de populație, și anume: (0-14 ani), (15-64 ani), (65 ani și peste). Vezi: V. Sora, I. Hristache, C. Mihăescu, *Demografie și statistică socială*, Ed. Economică, București, 1996, pp. 91-92.

a) În funcție de natura elementelor, colectivitățile pot fi formate din elemente cu un conținut material sau din elemente cu un conținut imaterial. Ființele și lucrurile formează *colectivități de stări* și se definesc la un moment anume, pe când evenimentele, faptele formează *colectivități de mișcări*, care se produc în mod continuu, definindu-se pe o perioadă de timp.

De exemplu:

- ansambluri de ființe (populația unei țări la un recensământ);
- stocuri de obiecte concrete (parcul de autoturisme românești la 1 ianuarie 2003);
- ansambluri de evenimente (cererile de angajare depuse la firma „A" în decursul unui an calendaristic);
- ansambluri de elemente neconcrete (opiniile electoratului înregistrate printr-o anchetă).

Volumul unei colectivități. Volumul unei colectivități (talia sau efectivul) reprezintă ansamblul indivizilor definiți prin omogenitate statistică (în timp, în spațiu și din punct de vedere calitativ).

Volumul unei colectivități se stabilește în mod diferențiat, în raport cu natura elementelor componente, astfel:

- în cazul colectivităților de mișcări (evenimente, fapte), volumul se află prin înregistrarea unităților statistice pe măsura apariției lor și prin însumarea acestora pentru un interval de timp ales;
- în cazul unei colectivități de stări, volumul se află prin numărarea elementelor componente existente la un moment dat.

Volumul colectivităților de stări, definite la diferite momente, poate fi aflat și cu ajutorul volumului colectivităților de mișcări corespunzătoare (vezi figura 2.2).



Figura 2.2 *Volumul colectivităților de stări la momentele T_0 și T_1*

De exemplu, volumul colectivității „populația unei zone" se poate afla pentru diferite momente de timp ținând seama de evenimentele demografice (nașteri, decese, intrări și ieșiri migratorii) produse în perioada dintre momente.

Astfel, *Populația la un moment dat* = Populația la un moment anterior momentului de calcul + Sporul natural, ca rezultat al evenimentelor demografice (nașteri și decese) produse în zonă, în perioada dintre momentul de calcul și momentul ales, anterior acestuia + Sporul migrator (intrări și ieșiri migratorii) în aceeași perioadă considerată.

b) În funcție de numărul elementelor componente, pot fi: *colectivități totale*, care cuprind totalitatea elementelor componente, și *colectivități parțiale* (de selecție sau eșantioane), care cuprind un număr reprezentativ de unități extrase dintr-o colectivitate totală. Practica înregistrării unui eșantion reprezentativ în locul colectivității totale poate fi impusă fie din motive de economicitate, fie din pricina faptului că nu avem acces la întreaga populație, fie pentru că prin înregistrare elementele colectivității s-ar distruge.

2.5.2 Unități statistice

Unitățile statistice reprezintă elementele componente ale unei colectivități statistice. De exemplu, unitățile statistice ale populației unei țări sunt indivizii. Unitățile statistice sunt *elemente de observare, măsurare și înregistrare*, adică prin ele se observă, măsoară și înregistrează o populație. Unele unități pot fi concrete, altele pot fi abstracte și nu servesc decât la individualizarea observațiilor.

Unitățile statistice trebuie să fie clar definite, cerință impusă de necesitatea identificării lor corecte pe teren, altfel s-ar crea confuzii în interpretare și, ca urmare, s-ar obține date neautentice.

De exemplu, în cazul populației unei localități, exprimată în număr de locuitori, unitățile statistice (indivizii) ar apărea la prima vedere perfect definite. Însă, ținând cont de militarii în termen sau de numărul studenților ce provin din alte localități, noțiunea de „locuitor" va fi de natură diferită. Situația este și mai dificilă în cazul unităților statistice definite în funcție de modul lor de organizare, cum ar fi, de exemplu, familia. Este necesar așadar să se definească precis unitățile statistice, respectiv să se cunoască categoriile de unități statistice.

Clasificarea unităţilor statistice. De regulă, se folosesc două criterii de clasificare a unităţilor statistice, şi anume: gradul de complexitate şi rolul pe care îl au în procesul înregistrării statistice.

a) După gradul de complexitate sau componenţa lor, pot fi *unităţi statistice simple* şi *unităţi statistice complexe.* Cele simple sunt formate dintr-un singur element (individul, de exemplu) şi depind de starea lor naturală de existenţă, pe când unităţile complexe (familia, de exemplu) sunt formate din două sau mai multe unităţi simple şi depind de modul lor de organizare.

b) După rolul lor în procesul înregistrării statistice, pot fi *unităţi statistice active* şi *unităţi statistice pasive.* Cele active transmit direct date statistice atât asupra lor, cât şi asupra unităţilor statistice pe care le reprezintă. De exemplu, capul de familie transmite, într-un recensământ, atât date cu privire la propria persoană, cât şi date cu privire la minorii pe care-i reprezintă. Unităţile pasive sunt unităţile despre care se transmit date.

2.5.3 Variabile statistice

Într-un studiu statistic, pe diferite trepte ale cercetării se pot întâlni trei tipuri de variabile: empirice, teoretice şi de selecţie. Corespunzător celor trei tipuri de variabile, se pot construi trei tipuri de distribuţii: distribuţii empirice sau statistice, distribuţii teoretice şi distribuţii de selecţie.

Variabile şi distribuţii statistice. Variabilele statistice sunt cunoscute în literatura de specialitate şi sub denumirea de *caracteristici statistice* şi reprezintă *şiruri de valori reale* înregistrate la nivelul unităţilor statistice ale unei colectivităţi bine definite. Exprimă însuşiri, trăsături esenţiale purtate de unităţile statistice ale unei colectivităţi, adică dimensiunile prin care se observă, respectiv se măsoară, cuantifică şi înregistrează fiecare unitate din colectivitate.

De exemplu, în cazul populaţiei umane, fiecare persoană este caracterizată prin sex, vârstă, stare civilă, naţionalitate, religie, ocupaţie etc.

Valorile unei variabile statistice se numesc *variante* ale variabilei şi se obţin prin observarea unităţilor unei colectivităţi statistice, la un moment dat sau într-un interval de timp.

Variantele unei variabile statistice pot diferi de la o unitate statistică la alta sau de la un grup de unităţi statistice la altul. Variaţia nivelului unei variabile statistice de la o unitate la alta se produce sub acţiunea unei multitudini de

factori, cu intensităţi şi sensuri de influenţă diferite, şi dă variabilelor statistice *caracterul de variabilă aleatorie.*

Şirul variantelor unei variabile cu frecvenţele de apariţie asociate formează o *distribuţie statistică*, numită şi *distribuţie empirică* sau *distribuţie observată* ori *distribuţie de frecvenţă*.

Distribuţiile statistice se diferenţiază între ele în funcţie de numărul variabilelor definitorii (distribuţii statistice unidimensionale, bidimensionale şi multidimensionale), de natura şi modul de măsurare a variabilelor.

Variabile şi distribuţii aleatorii. O variabilă aleatorie este un şir de valori abstracte. Este numită şi *variabilă teoretică* şi are un caracter stochastic, variantele variabilei depinzând de un sistem complex de evenimente întâmplătoare. Probabilităţile de realizare a variantelor sunt cu atât mai mari cu cât şansa de influenţă a factorilor determinanţi este mai mare.

Şirul variantelor unei variabile aleatorii cu probabilităţile de apariţie corespunzătoare formează o *distribuţie aleatorie*, numită şi *distribuţie teoretică* sau *distribuţie de probabilitate.*

Variabile şi distribuţii de selecţie. Variabilele de selecţie se întâlnesc în cazul cercetării prin sondaj. Pentru un volum de selecţie n sunt numite *variabile de selecţie* variabilele aleatorii X_1, X_2, ... X_n, independente stochastic în ansamblu şi identic distribuite cu variabila X a populaţiei. Orice funcţie de variabile de selecţie este numită *statistică*, de exemplu, media, varianţa unei colectivităţi. O *distribuţie de selecţie* este o distribuţie a unei statistici, de exemplu, distribuţia mediei de selecţie.

Clasificarea variabilelor statistice. În clasificarea variabilelor statistice, se consideră, de regulă, următoarele criterii: importanţa lor în procesul cunoaşterii colectivităţii, natura, modul de exprimare, forma de manifestare.

a) După importanţa lor, variabilele pot fi esenţiale şi neesenţiale. *Variabilele esenţiale* exprimă natura internă a fenomenului, de exemplu, sexul persoanelor, şi sunt purtate de toate unităţile colectivităţii. Variabilele esenţiale diferenţiază colectivităţile unele de altele.

Variabilele neesenţiale au caracter întâmplător şi pot fi purtate numai de unele unităţi din colectivitate, de exemplu, vechimea în calitate de membru al unui club.

Unităţile statistice ale unei colectivităţi posedă un număr foarte mare de caracteristici, de exemplu:

- sexul, vârsta, starea civilă, ocupația etc., în cazul indivizilor unei populații umane;
- tipul, culoarea, vechimea, puterea motorului etc., în cazul mașinilor dintr-un parc de autoturisme.

Observație! Într-un studiu statistic, din multitudinea caracteristicilor pe care le posedă fiecare unitate se rețin numai acelea care prezintă interes pentru cercetarea întreprinsă și sunt cuprinse în programul statistic de înregistrare.

b) După natura lor, pot fi: variabile calitative, variabile de timp și variabile de spațiu.

Variabilele de timp desemnează apartenența unităților la un moment sau la un interval de timp.

Variabilele de spațiu, numite și *teritoriale*, exprimă teritoriul în care există și se manifestă unitățile colectivității.

Variabilele calitative exprimă esența, natura unităților.

c) După modul de exprimare, pot fi: *variabile numerice* și *variabile nenumerice*.

Variabilele numerice, numite și *cantitative*, sunt fie numărabile, fie măsurabile, respectiv cu variație discontinuă – numite *variabile discrete*, sau cu variație continuă – numite *variabile continue*. Valorile unei variabile numerice se stabilesc prin numărare, măsurare, calcul și pot fi reprezentate pe o scală interval sau pe o scală raport.

Variabila discretă este caracterizată prin „întreruperea" valorilor pe care le poate lua această variabilă. Variabila discretă ia *valori numărabile* (de exemplu, numărul de piese produse zilnic de un muncitor, numărul de copii pe o familie, producția de autoturisme). Se exprimă, de regulă, în numere întregi, nonnegative.

Variabila continuă este o variabilă numerică măsurabilă, ale cărei valori sunt divizibile la infinit și pot fi grupate în k intervale. Exprimă *dimensiuni măsurabile* (de exemplu, puterea mașinilor, salariul angajaților unei firme, lungimea unei piese, greutatea unui produs). O variabilă continuă presupune alegerea unităților de măsură și a preciziei dorite pentru rezultate. Fiecărui element dintr-o colectivitate îi corespunde un nivel al variabilei exprimat numeric, în unități de măsură corespunzătoare (de exemplu, puterea mașinilor în C.P., salariul în lei etc.).

În cazul unei variabile continue, valorile observate pentru fiecare unitate din colectivitatea studiată (pentru fiecare salariat, de exemplu) fac parte din

intervale de valori. Intervalele (grupele, clasele de valori) pot fi de mărimi egale sau inegale, închise sau deschise.

Observație! În realitate, și observațiile pentru o variabilă continuă sunt discontinue, rotunjite la câteva cifre (discontinue în raport cu unitatea de măsură folosită: lei, metri, kilograme etc.). De asemenea, se poate observa că atunci când numărul valorilor unei variabile discrete este mare (de exemplu, numărul de clienți care intră într-un magazin pe parcursul unei zile), aceste valori pot fi prezentate, pe intervale, ca în cazul unei variabile continue.

Variabilele nenumerice, cunoscute și sub denumirea de variabile atributive, categoriale, nominale, sunt caracteristici ale căror valori (modalități) de manifestare sunt *exprimate atributiv, în cuvinte* (de exemplu, sexul, naționalitatea, culoarea). Când numărul variantelor unei caracteristici atributive este mare, acestea fac obiectul *nomenclatoarelor statistice*, de exemplu, nomenclatorul meseriilor.

Modalitățile unei caracteristici atributive pot fi reprezentate pe o *scară nominală*, de exemplu, meseriile, sau pe o *scală ordinală*, de exemplu, calificarea profesională (cu modalitățile calificat, semicalificat, necalificat). Corespunzător, variabilele atributive pot fi:

- variabile nominale;
- variabile nominal-ordinale.

Observație! Distincția între o variabilă exprimată cantitativ și o variabilă exprimată atributiv este câteodată convențională. Adică, unei caracteristici numerice i se pot asocia modalități, atribute. De exemplu, în cazul tipurilor după vârstă, se pot asocia atributele astfel: tânăr – până la 20 de ani, adult – 20-60 de ani și în vârstă – peste 60 de ani. Dar operația inversă, adică atribuirea unor valori numerice unei caracteristici exprimate atributiv, nu este posibilă.

d) După modul de manifestare la nivelul unităților simple ale unei colectivități, se disting variabile nealternative și variabile alternative.

Variabilele nealternative pot lua valori diferite pentru fiecare unitate statistică sau grupă de unități statistice din colectivitatea observată.

Variabilele alternative au caracter dichotomic, adică nu pot lua decât două valori. Mai sunt denumite și *caracteristici binomiale* sau *binare*. De exemplu, un produs poate fi rebut sau nonrebut, bun sau rău, un candidat poate fi admis sau respins.

Valorile individuale pentru o caracteristică alternativă pot fi:
- *DA*, exprimând prezenţa caracteristicii şi are asociat *codul numeric 1*;
- *NU*, exprimând absenţa caracteristicii şi are asociat *codul numeric 0*.

Codificare. Codificarea variabilelor statistice presupune atribuirea de coduri numerice fiecărei variante sau fiecărui interval, respectiv atribut (modalitate). Codurile sunt numere asociate pentru fiecare clasă, respectiv pentru fiecare modalitate (atribut). Reprezintă identificatori, acordarea lor este pur convenţională, deci codurile nu se supun operaţiilor de prelucrare statistică.

2.5.4 Cuantificarea fenomenelor. Tipuri de scală

Modul de măsurare a fenomenelor şi proceselor economico-sociale, naturale etc. este diferenţiat, deoarece acestea se deosebesc între ele prin mărime sau prin formă, prin frecvenţă sau intensitate de manifestare, prin coeziune sau prin interdependenţe. Ca urmare, unele fenomene pot fi *direct măsurabile cantitativ* – cazul elementelor fizice –, altele pot fi *măsurabile cu aproximaţie*[6] – cazul elementelor sociale, care, prin natura lor, sunt mai dificil de măsurat comparativ cu elementele fizice. De exemplu, distanţa socială este mai dificil de măsurat decât distanţa geografică sau opinia unei persoane comparativ cu înălţimea sa.

În vederea caracterizării colectivităţilor statistice este necesară obţinerea unor forme comparabile ale fenomenelor. Aceasta devine posibilă prin cuantificare.

Cuantificare

Cuantificarea fenomenelor economico-sociale implică *un proces complex de izolare*, *măsurare* în forme comparabile şi *înregistrare* a elementelor unei colectivităţi prin caracteristicile cuprinse în programul înregistrării. Operaţia de cuantificare este o operaţie specific statistică şi *presupune un set de reguli de atribuire a unei valori unităţilor statistice* ale unei colectivităţi observate după o caracteristică. Valorile atribuite pot fi sub formă de cifre sau de simboluri. Diferenţierea valorilor se face prin intermediul unui instrument de măsură denumit *scală*.

6. O analiză aprofundată a fenomenelor din punctul de vedere al posibilităţii de măsurare găsim în M. Dauverger, „Tehnicile matematice şi grafice în ştiinţele sociale", în *Teorie şi metodă în ştiinţele sociale*, vol. II, Editura Politică, Bucureşti, 1956, p. 125.

Tipuri de scală

Scala poate fi considerată un *continuum de cifre sau de simboluri* plasate ierarhic, de la inferior la superior. În literatura de specialitate şi în practica statistică sunt folosite diferite scale, şi anume: scala nominală, scala ordinală, scala interval şi scala raport.

Scala nominală are o singură proprietate – *identitatea.* Aceasta exprimă apartenenţa elementelor la o categorie. Ca urmare, măsurarea cu ajutorul scalei nominale presupune *existenţa unei colectivităţi împărţite în categorii (modalităţi) independente* şi constă în *acordarea de numere sau simboluri fiecărei categorii* în care se diferenţiază unităţile colectivităţii observate. Scalarea cu ajutorul scalei nominale cere ca:

- tuturor indivizilor unei categorii să li se atribuie aceeaşi valoare;
- doi indivizi care aparţin la două categorii distincte să aibă valori diferite.

De exemplu, pentru variabila dichotomică „sex" se pot da două atribute: masculin şi feminin sau, cu ajutorul codurilor, două numere: 1 şi 2. Ordinea numerelor sau simbolurilor atribuite drept coduri categoriilor este oarecare, între ele există un *raport de echivalenţă.*

Scala ordinală sau *scala cu ranguri* are, pe lângă *proprietatea de identitate* specifică scalei nominale, şi *proprietatea de ordine*, care permite să se claseze elementele observate conform unei ordini, preferinţe. Măsurarea cu ajutorul scalei ordinale *este folosită în cazul variabilelor atributive,* când între categoriile de unităţi ale colectivităţii există un *raport de preferinţă*, desemnat prin „>", care permite măsurarea cu ranguri.

Cel mai frecvent caz de măsurare ordinală este nota la examene: 1, 2, 3, ..., 10. Indiferent de sistemul de notare (de la 1 la 10 sau de la 1 la 100), 10 sau 100 are aceeaşi interpretare, reprezentând categoria „cel mai bun". Diferenţele între două valori ale unei variabile ordinale nu au sens.

Scala interval adaugă la cele două proprietăţi subliniate anterior, *identitate şi ordine*, o a treia proprietate – *intervalul între numere are un sens, fapt ce permite să se compare diferenţele dintre numere.*

Măsurarea cu ajutorul scalei interval este utilizată în cazul variabilelor cantitative şi presupune atribuirea de valori numerice unităţilor colectivităţii în funcţie de sensul diferenţei de mărime a caracteristicii observate.

Diferenţa dintre două valori ale unei variabile este semnificativă, pentru scala interval *fiind caracteristic raportul dintre două puncte de scală.* Acesta

este independent de punctul de origine ales şi de unităţile de măsură folosite, fapt ce face posibilă trecerea de la un sistem de măsurare la altul.

Exemplul clasic îl reprezintă măsurarea temperaturii în sistemul Celsius şi în sistemul Fahrenheit – caz în care, schimbând zeroul convenţional şi valorile temperaturii, raportul dintre două modificări de temperatură rămâne acelaşi.

Scala raport este folosită tot *pentru variabile cantitative* şi are *aceleaşi proprietăţi ca scala interval* şi, în plus, *posedă un zero absolut*. Măsurarea cu ajutorul scalei raport presupune respectarea raportului specific unei scale interval, plus considerarea unei *origini reale, fixe, ca punct de referinţă.*

Diferenţele şi raporturile dintre două valori ale unei variabile au un sens precis. De asemenea, raportul dintre oricare două valori ale scalei este independent de unitatea de măsură folosită. Scala raport este utilizată pentru măsurarea valorilor pentru numeroase variabile, cum ar fi, de exemplu, dimensiunile fizice (înălţime, greutate, distanţă etc.), preţul, viteza etc.

În folosirea practică a scalei raport apar dificultăţi în cazul măsurării valorilor unei variabile continue, ca urmare a limitelor de precizie ale instrumentelor de măsură utilizate.

Scala de intensitate este folosită pentru măsurarea şi compararea opiniilor, a comportamentelor. S-au elaborat *scale specifice de intensitate*. O contribuţie deosebită în cuantificarea opiniilor a adus-o Ilse Krasemann[7], prin *scala de opinie*, pe care o prezentăm în figura 2.3.

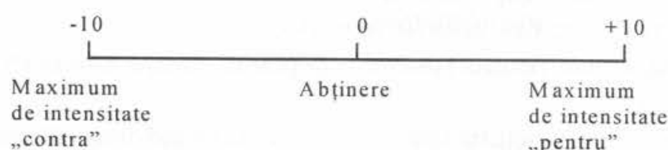


Figura 2.3 *Scală de opinie*

Scala de opinie este o scală „cvasimetrică". Se caracterizează prin punctul *0* care exprimă inexistenţa opiniei şi un număr de puncte, în sens negativ şi pozitiv, faţă de zero, cu ajutorul cărora se exprimă şi se măsoară gradele de intensitate a opiniei subiecţilor anchetaţi.

7. Ilse Krasemann, „Unele observaţii cu privire la cuantificarea fenomenelor sociale", *ibidem*, vol. X, p. 103.

În cercetările de marketing[8], scala de opinie este cunoscută, de regulă, sub denumirea de *scală de rating*. Pe o astfel de scală sunt fixate de la 4 până la 10 gradații pentru a facilita ierarhizarea răspunsurilor.

De exemplu, pentru un item (enumerare) în 5 trepte, gradațiile pot fi formulate astfel:

- *coincide perfect*;
- *coincide parțial*;
- *coincide într-o oarecare măsură*;
- *coincide mai puțin*;
- *nu coincide deloc*.

În vederea prelucrării, răspunsurile se cuantifică, adică fiecărei gradații i se asociază o cifră, de aici rezultând și proprietatea de *scală „cvasimetrică"* pentru scala de opinie.

2.6 Notații

Procesul cunoașterii statistice, descriptiv și inferențial, implică planuri de abordare diferite, și anume: *distribuții empirice, distribuții teoretice* și *distribuții de selecție*. Pentru a evita confuziile este necesară folosirea unor notații distincte pentru fiecare plan. În lucrarea de față vom ține seama de acest principiu, respectând, pe cât posibil, notațiile cel mai des întâlnite în literatura de specialitate, și anume:

- *litere din alfabetul latin* pentru *valorile observate*, fie prin anchete statistice, fie prin experiment;
- *litere grecești* pentru *valorile teoretice*;
- *litere majuscule* pentru *variabile* și pentru *funcții cumulative* (funcții de repartiție);
- *litere minuscule* pentru *valorile particulare* ale unei variabile și pentru *funcții noncumulative* (funcții de densitate).

Astfel, în concordanță cu precizările anterioare, notăm prin literă majusculă, de exemplu, X, o *variabilă statistică*, reprezentând *o funcție cu valori reale definită pe mulțimea de bază* Ω:

$X: \Omega \to R$

O variantă a variabilei purtată de un grup n_i de unități se notează cu minusculă, de exemplu, x_i.

8. Manfred Bruhn, *Marketing*, Editura Economică, București, 1999, p. 114.

Observaţie! Identificatorul „*i*" este folosit pentru a desemna poziţia unui individ într-un şir, într-o listă, respectiv pentru a *desemna valoarea variabilei înregistrate la nivelul unităţii sau individului „i" din şir.*

În cazul unei variabile discrete, valorile unei variabile sunt:
$x_1, x_2, ..., x_n$, unde ($x_1 < x_2 < ... < x_n$).

În cazul unei variabile continue, admitem intervalele de valori:
$(x_0, x_1), (x_1, x_2), ..., (x_{i-1}, x_i), ..., (x_{k-1}, x_k)$.

Oricare interval (x_{i-1}, x_i) poate fi notat: $j_i = (x_{i-1}, x_i)$, unde (x_{i-1}, x_i) este un segment de dreaptă real.

Domeniul de variaţie sau *amplitudinea de variaţie a variabilei* este reprezentat de mulţimea valorilor posibile ale unei variabile, $X: (x_1, x_2, ..., x_n)$. Se notează cu A_x şi se calculează ca diferenţă între nivelul maxim (x_{max}) şi nivelul minim (x_{min}), înregistrate pentru unităţile colectivităţii observate, după relaţia:

$$A_x = x_{max} - x_{min}.$$

Când variabila X ia valoarea particulară x_i, spunem că are loc *evenimentul* $X = x_i$ *cu probabilitatea* $P(X = x_i) = f(x_i)$. Adică, unei variabile aleatorii X i se asociază o funcţie bine definită, $f(x)$, care indică probabilitatea de apariţie a unei valori posibile, x_i.

Observaţie! Statistic, probabilitatea de apariţie a unei valori x_i poate fi aproximată prin frecvenţa relativă, f_i, corespunzătoare acestei valori, respectiv prin ponderea efectivelor care poartă nivelul x_i în totalul unităţilor colectivităţii distribuite după variabila X, şi anume: $f_i = \frac{n_i}{n}$.

Distribuţia de frecvenţă a variabilei statistice X este reprezentată de mulţimea perechilor (x_i, f_i), $i = \overline{1,m}$, respectiv $(j_i, f_i), i = \overline{1,k}$.

Sintetic, o distribuţie statistică se notează astfel:

$$X: \begin{pmatrix} x_i \\ n_i \end{pmatrix} \text{ sau } X: \begin{pmatrix} x_i \\ f_i \end{pmatrix}, \text{ cu } i = \overline{1,m},\ 1 \le m < n, \text{ respectiv}$$

$$X: \begin{pmatrix} j_i \\ n_i \end{pmatrix} \text{ sau } X: \begin{pmatrix} j_i \\ f_i \end{pmatrix} \text{ cu } i = \overline{1,k}.$$

unde:

x_i – valori individuale ale caracteristicii X;

$j_i = (x_{i-1}, x_i)$ – intervale de valori ale unei variabile continue;

n_i – efectivul, frecvența absolută;

f_i – frecvența relativă corespunzătoare valorii $X = x_i$, respectiv $X = j_i$.

CAPITOLUL 3

PREGĂTIREA, SISTEMATIZAREA ŞI PREZENTAREA DATELOR ÎN SPSS

- **Definirea şi introducerea datelor**
- **Divizarea unui fişier**
- **Sistematizarea şi prezentarea datelor în SPSS**
- **Transformarea datelor**
- **Modificarea unui tabel în SPSS**

3.1 Definirea şi introducerea datelor

Orice analiză statistică a datelor în SPSS începe cu pregătirea setului de date. Acest proces presupune prezentarea datelor într-un format care să permită organizarea şi efectuarea analizei lor. Atingerea acestui obiectiv implică definirea şi introducerea datelor, operaţii care se efectuează folosind foile *Data View* şi *Variable View* din fereastra *Data Editor*.

Ilustrarea şi exemplele din această carte au la bază fişierele *tapestry.sav*[1] şi *dez_reg.sav*. Crearea fişierului de date, pentru primul caz, are la bază un *chestionar* (vezi Anexa 1) administrat unui eşantion de 400 de persoane, iar pentru cel de al doilea caz, Anuarul statistic al României, 2002.

3.1.1 Definirea atributelor unei variabile

Definirea atributelor unei variabile este prima operaţie din procesul de pregătire a setului de date. Presupune precizarea atributelor unei variabile: *numele variabilei, tipul, lungimea (numărul de caractere), numărul de zecimale (pentru cele numerice sau asociate celor numerice), eticheta, valorile etichetei, valorile lipsă, alinierea şi modalităţile de măsurare a variabilei (scală, ordinal sau nominal)*. Variabilele se definesc în coloanele foii *Variable View* din fereastra *Data Editor* (vezi figura 3.1).

tapestry.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
52	sexul	Numeric	8	0	sexul persoan	{1, masculin}...	None	8	Right	Nominal
53	varsta	Numeric	8	0	varsta persoan	None	None	8	Right	Scale
54	statut	Numeric	8	0	statutul socio-	{1, angajat per	None	8	Right	Ordinal
55	venitl	Numeric	8	0	venitul lunar al	{1, mai putin d	None	8	Right	Ordinal
56	jud_tara	Numeric	8	0	judetul in care	{1, Brasov}...	None	8	Left	Nominal

Data View / Variable View

SPSS Processor is ready

Figura 3.1 *Fereastra* Data Editor – *Foaia* Variable View

1. *tapestry.sav* – baza de date cu privire la eşantionul de *pelerini la Sfânta Cuvioasă Parascheva, Iaşi, octombrie 2002* (Research Project TAPESTRY: Travel Awareness Publicity and Education Supporting a Sustainable Transport Strategy in Europe).

1. Numele variabilei

Numele variabilei se editează în coloana *Name*, ținând cont de câteva restricții:

- să fie unic;
- să aibă cel mult 8 caractere;
- primul caracter să fie o literă;
- poate să conțină litere, cifre (inclusiv o perioadă) și simbolurile @, #, _, $;
- să nu conțină spații sau simboluri speciale folosite în SPSS;
- ultimul caracter să nu fie „_" (*underscore* – liniuța de subliniere);
- să nu se termine cu o perioadă.

2. Tipul variabilei

Definirea acestui atribut se realizează în coloana *Type* din foaia *Variable View* din fereastra *Data Editor*. Variabilele pot fi de mai multe tipuri: numerice (*Numeric*, *Comma*, *Dot*, *Scientific notation*), alfanumerice (*String*) etc. (vezi figura 3.2). Variabilele introduse sub formă alfanumerică (de exemplu, sexul persoanelor, cu atributele: M – pentru masculin și F – pentru feminin), cel mai adesea, trebuie codificate (se asociază un număr pentru fiecare atribut), în scopul efectuării analizelor ulterioare. Prin urmare, se recomandă editarea lor numerică de la început.

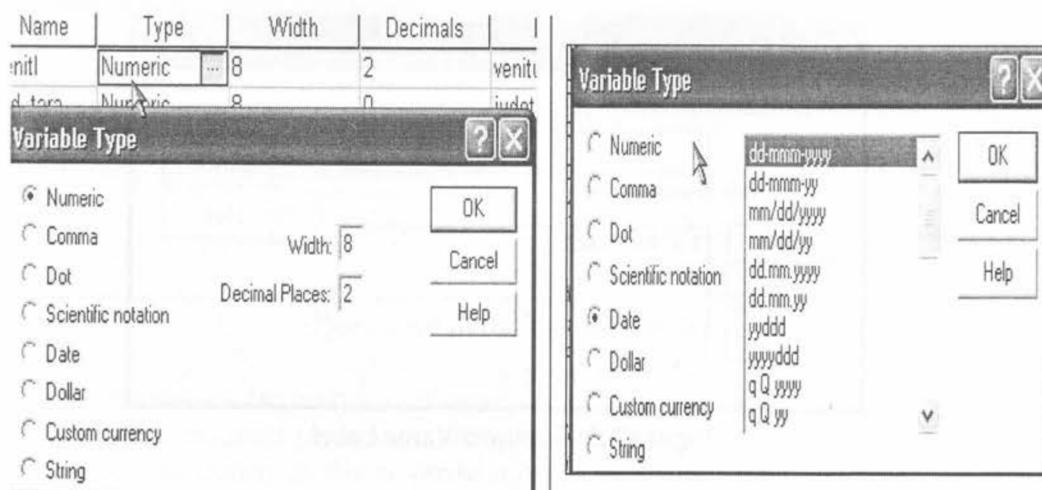


Figura 3.2 *Fereastra* Variable Type

Pentru datele de tip *Numeric*, *Comma*, *Dot* și *Scientific notation*, se pot introduce numere întregi și zecimale, dar vor fi afișate sub formă zecimală

numai dacă precizăm numărul de zecimale în caseta *Decimal Places* sau direct în coloana *Decimals*.

Pentru variabilele de tip *Date*, *Dollar* şi *Custom currency*, sunt afişate liste cu formate specifice din care îl alegem pe cel dorit (vezi figura 3.2).

3. Eticheta variabilei

Numele variabilelor este limitat la 8 caractere, dar se poate preciza un nume explicit, numit etichetă, care să fie afişat în fereastra de rezultate *Output* (de exemplu, sexul persoanei). Pentru aceasta, în coloana *Label* se poate edita un nume folosind până la 256 de caractere.

4. Valorile etichetei

Când variabila este categorială (nominală), se precizează valorile luate de variabilă şi etichetele corespunzătoare acestora, în fereastra *Value Label* (vezi figura 3.3). De exemplu, pentru sexul persoanei se scrie 1 în *Value* şi *Masculin* în *Value Label*. Se acţionează butonul de comandă *ADD* şi, în mod analog, se adaugă noi valori (de exemplu, 2 în *Value* şi *Feminin* în *Value Label*).

Pentru modificarea unor valori de etichetă, se foloseşte butonul de comandă *Change*, iar pentru ştergerea lor, butonul de comandă *Remove*. Butonul de comandă *OK* este acţionat după ce au fost adăugate, şterse sau modificate toate valorile dorite ale variabilei.

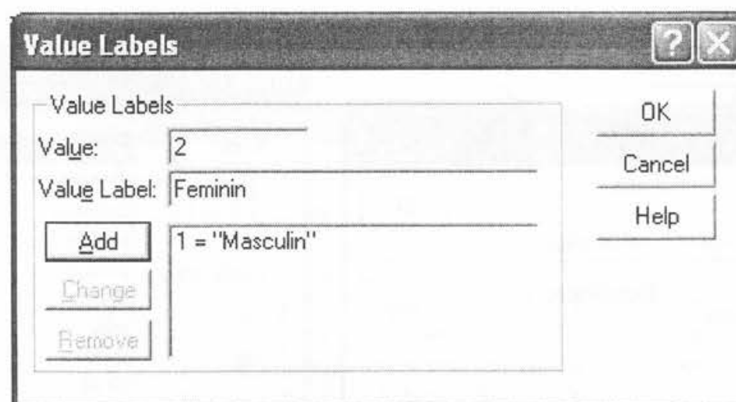


Figura 3.3 *Fereastra* Value Label

5. Precizarea valorilor lipsă

În SPSS se pot preciza două tipuri de valori lipsă: *system-missing values* (pentru variabile sistem) şi *user-missing values* (pentru variabile definite de utilizator). Valorile lipsă trebuie precizate pentru a asigura acurateţe

rezultatelor. Ele apar când un set de date este incomplet din diferite cauze: fie chestionarul administrat este prea lung şi chestionaţii nu mai au răbdare să completeze răspunsul la toate întrebările, fie se fac erori de omitere în procesul de introducere a datelor, fie completările sunt ilizibile etc. În astfel de situaţii, pentru un subiect particular, când răspunsul nu întruneşte criteriile pentru a fi considerat valid în procesul de editare a datelor în SPSS, în celula corespunzătoare se scrie valoarea 9 sau 99, respectiv 999, în funcţie de numărul de cifre din răspunsul normal. De asemenea, în practica anchetelor prin sondaj, pentru astfel de răspunsuri invalide se folosesc codurile:

97 – pentru „nonrăspuns";

98 – pentru „neaplicabil";

99 – pentru „răspuns ilizibil".

Pentru a recunoaşte valorile lipsă, acestea trebuie definite. Când nu lipsesc valori, se alege butonul de opţiuni *No missing values*.

În procesul de analiză a datelor, se pot preciza ca valori lipsă şi valorile aberante. Acestea se introduc, de regulă, în zonele de editare subordonate butonului de opţiuni *Discrete missing value* (vezi figura 3.4).

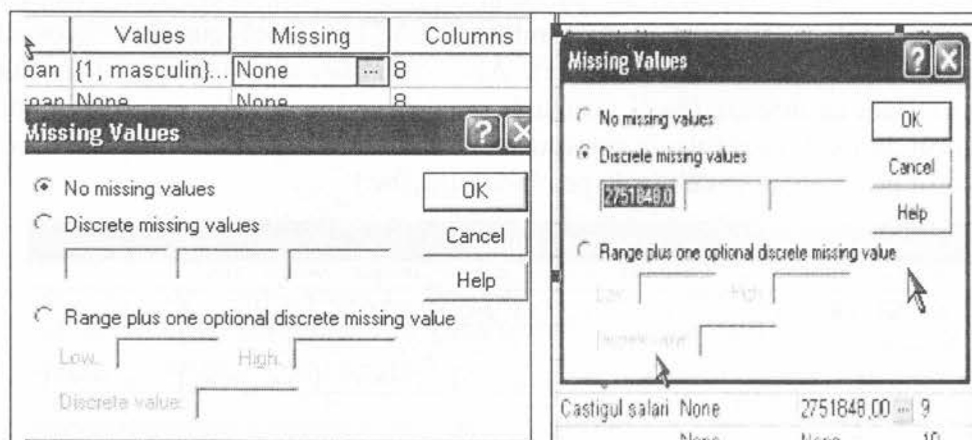


Figura 3.4 *Definirea valorilor lipsă*

6. Definirea formatului coloanei

Formatul coloanei presupune precizarea numărului de caractere (automat, în *Columns* este definit 8, dar se poate schimba înlocuind 8 cu valoarea dorită) şi alinierea valorilor în coloană (*Left* – stânga, *Right* – dreapta sau *Center* – centru). De asemenea, se poate alege şi sistemul de măsurare (*Scale*, *Ordinal* sau *Nominal*). Aceste opţiuni se aleg din listele afişate în coloanele *Align* şi *Measure*, din foaia *Variable View* (vezi figura 3.5).

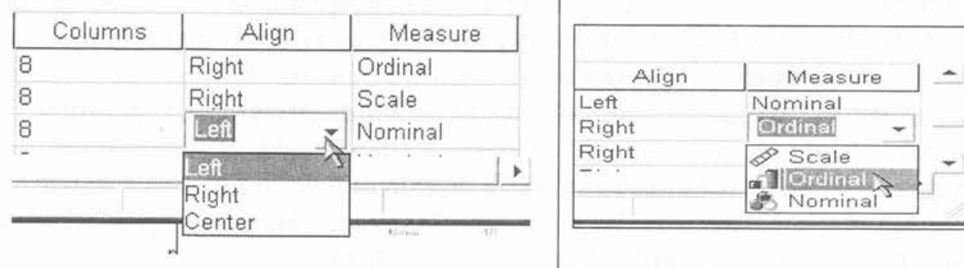


Figura 3.5 *Definirea formatului coloanei*

3.1.2 Introducerea datelor

Datele se introduc în celulele foii *Data View* din fereastra *Data Editor*, deschisă prin comanda *New Data*, din meniul *File*. Fiecare rând reprezintă un subiect, fiecare coloană reprezintă o variabilă. Introducerea este uşoară şi se realizează prin scrierea (editarea) unui număr sau a unui text în celula curentă (cea în care este cursorul şi are chenarul îngroşat – vezi figura 3.6). Pentru a introduce date, succesiv, în mai multe celule, se mută cursorul folosind mouse-ul (prin clic în celula dorită). Alte modalităţi de mutare a cursorului oferă tastele de direcţie (de la tastatură, tastele cu săgeţi), tasta *Tab* (care mută cursorul pe rând, în celula de pe coloana următoare) sau tasta *Enter* (care mută cursorul pe coloană, în celula de pe rândul următor).

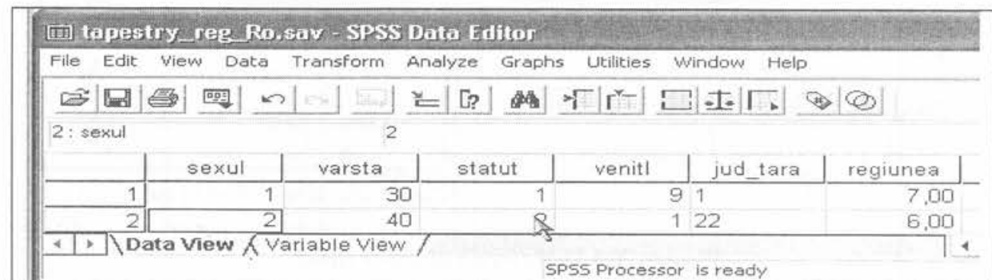


Figura 3.6 *Introducerea datelor*

De asemenea, pentru introducerea datelor se pot folosi comenzile de editare *Cut*, *Copy* şi *Paste* din meniul *Edit* sau meniul rapid (activat cu butonul din dreapta al mouse-ului).

3.1.3 Citirea atributelor variabilelor

Atributele unei variabile se pot citi alegând din meniul *Utilities* comanda *Variables* care deschide fereastra *Variables* (vezi figura 3.7). Această fereastră este structurată în două zone principale. În stânga, este afișată lista tuturor variabilelor, iar în dreapta apar informații despre variabila selectată (cea pe care este plasată bara de selecție), prin clic de mouse.

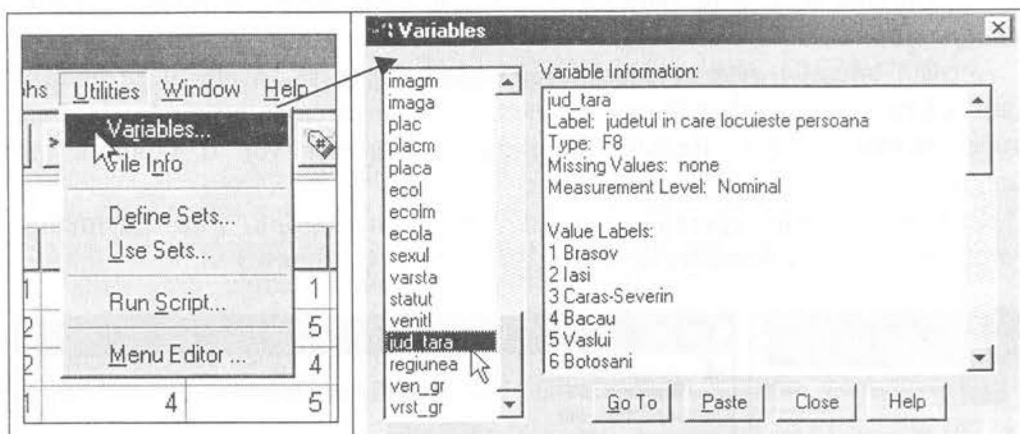


Figura 3.7 *Citirea atributelor unei variabile*

Fereastra *Variables* are și o serie de butoane de comandă. Cel mai folosit este *Go To* care asigură localizarea, pentru un anume subiect, a coloanei corespunzătoare unei variabile. Operația se realizează fie pentru a avea acces rapid la valorile variabilei, fie pentru definirea sau modificarea atributelor respectivei variabile (vezi figura 3.8).

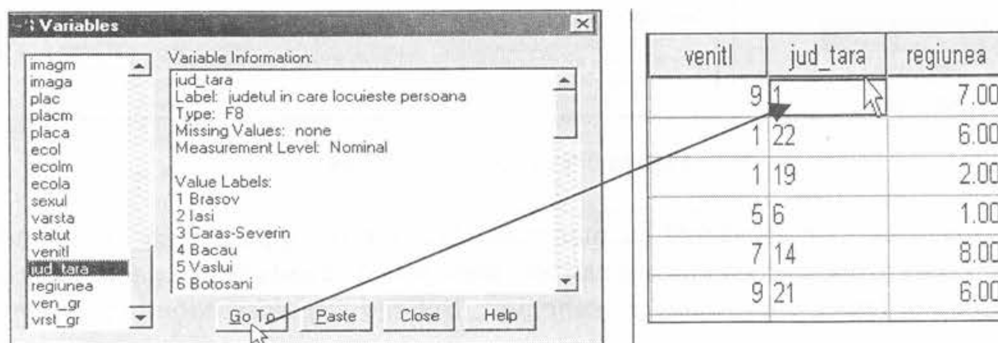


Figura 3.8 *Localizarea unei variabile*

3.2 Divizarea unui fişier

3.2.1 Divizarea unui fişier pe categorii de subiecţi, folosind comanda SPLIT FILE

Divizarea unui fişier de date pe categorii de subiecţi se face pe baza unei variabile categoriale prin care se definesc categoriile (grupurile). De exemplu, pentru a divide fişierul de date *tapestry.sav* în două, se poate folosi variabila *Sexul persoanei*.

Această operaţie este necesară atunci când dorim să se efectueze analiza statistică pe categorii de subiecţi, de exemplu, realizarea de teste pe cele două grupe: bărbaţi, femei. Rezultatele prelucrării datelor vor fi raportate pe categorii.

Realizarea acestei operaţii se bazează pe comanda *Split File* din meniul *Data*, care deschide fereastra de dialog *Split File* (vezi figura 3.9).

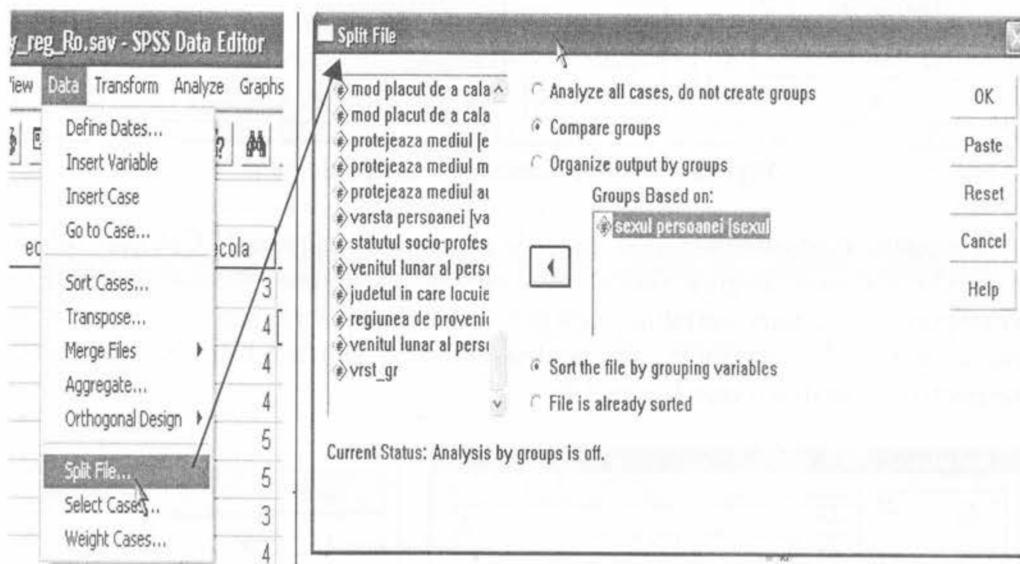


Figura 3.9 *Comanda* Split File

Fereastra este prevăzută cu mai multe butoane de opţiuni (care se exclud reciproc). *Compare groups* are ca efect prezentarea rezultatelor separat, pentru bărbaţi şi femei, fiecare într-un subraport. Butonul *Organize output by groups*

creează un raport cu toate informațiile pentru bărbați și un alt raport cu toate informațiile pentru femei.

Butonul de comandă *OK* realizează împărțirea/splitarea propriu-zisă pe grupe (în dreapta barei de informații apare mesajul *Split File On*). Ca urmare, rezultatele unei analize sunt afișate în *Output Viewer* ca raportate pe categorii de unități.

Pentru a reveni la forma inițială a fișierului de date se activează butonul de opțiuni *Analyze all cases, do not create groups* care asigură analizarea tuturor cazurilor, fără crearea de grupe.

3.2.2 Selectarea unor subiecți, folosind comanda SELECT CASES

Selectarea unor subiecți (de exemplu, numai persoanele până la 25 de ani) pe care dorim să îi analizăm în funcție de anumite caracteristici ale lor se poate realiza prin comanda *Select Cases*, din meniul *Data*. Această comandă deschide fereastra de dialog *Select Cases*. Se alege opțiunea *If condition is satisfied* și se acționează butonul de comandă *If* (vezi figura 3.10). Se deschide fereastra *Select Cases: If*, în care se introduce condiția de selecție/filtrare: *varsta < 25*.

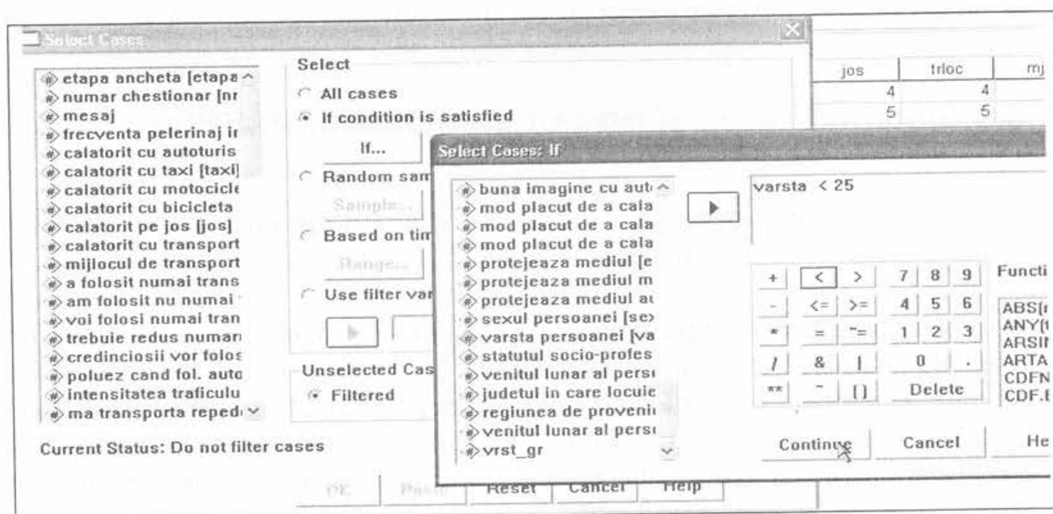
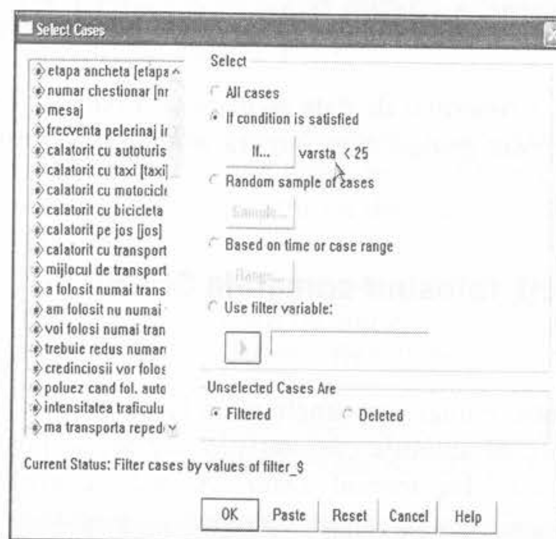


Figura 3.10 *Ferestrele* Select Cases *și* Select Cases: If

Butonul de comandă *Continue* determină revenirea la fereastra *Select Cases* în care se activează butonul de comandă *OK* pentru a se obține fișierul filtrat (vezi figura 3.11).



	varsta	statut
~~15~~	60	8
16	17	5
~~17~~	60	8
~~18~~	30	1
19	22	5
~~20~~	50	1
~~21~~	70	8
~~22~~	30	1

Figura 3.11 *Selectarea persoanelor în vârstă de până la 25 de ani*

În foaia *Data View* din fereastra *Data Editor*, cazurile anulate sunt tăiate printr-un slash (/). Aceste cazuri nu vor fi folosite în nici o raportare.

3.3 Sistematizarea şi prezentarea datelor în SPSS

Sistematizarea datelor este prima etapă a prelucrării datelor şi are ca obiectiv sumarizarea şi ordonarea datelor. Se poate realiza prin *centralizare şi grupare statistică*.

Prin centralizare se obțin *indicatori totalizatori* la nivelul unei populații, de exemplu: numărul locuitorilor unei țări la un moment de recensământ. Prin grupare, care poate fi tratată ca o centralizare pe grupe, se obțin *şiruri de date ordonate* după una sau mai multe variabile de grupare. Fiecare nivel al variabilei apare o singură dată, ordonat în sens crescător sau descrescător, la care se asociază frecvența de apariție corespunzătoare.

Şirurile de valori/categorii ordonate ale variabilei/variabilelor observate şi frecvențele asociate acestora formează *distribuții statistice.* Distribuțiile rezultate în urma sistematizării pot fi prezentate în *tabele statistice.*

Sistematizarea datelor după o variabilă presupune ordonarea valorilor variabilei observate, X, în sens crescător sau descrescător şi găsirea frecvenţei de apariţie corespunzătoare fiecărei valori sau grupe (clase) de valori. În urma sistematizării datelor pe baza unei variabile se obţin *distribuţii de frecvenţă univariate*, $X:(x_i,n_i)$, cu $i=\overline{1,m}$, care pot fi prezentate în *tabele de frecvenţă* – în cazul variabilelor numerice, respectiv în *tabele de contingenţă* – în cazul variabilelor nominale.

Sistematizarea datelor simultan după două sau mai multe variabile are ca rezultat obţinerea unei *distribuţii de frecvenţă bivariate*, $X,Y:(x_i,y_j,n_{ij})$, cu $i=\overline{1,m}$ şi $j=\overline{1,p}$, sau *multivariate*. Distribuţiile bivariate, rezultate în urma sistematizării, pot fi prezentate în *Crosstabs*, care pot fi tip *tabele de corelaţie* – cazul variabilelor numerice, respectiv *tabele de asociere* – cazul variabilelor nominale.

3.3.1 Demersul sistematizării datelor în SPSS

Sistematizarea datelor în SPSS poate fi realizată prin opţiunea *Frequencies* subordonată comenzii *Descriptive Statistics* din meniul *Analyze* (vezi figura 3.12). Activarea opţiunii *Frequencies* determină deschiderea ferestrei *Frequencies* (vezi figura 3.13).

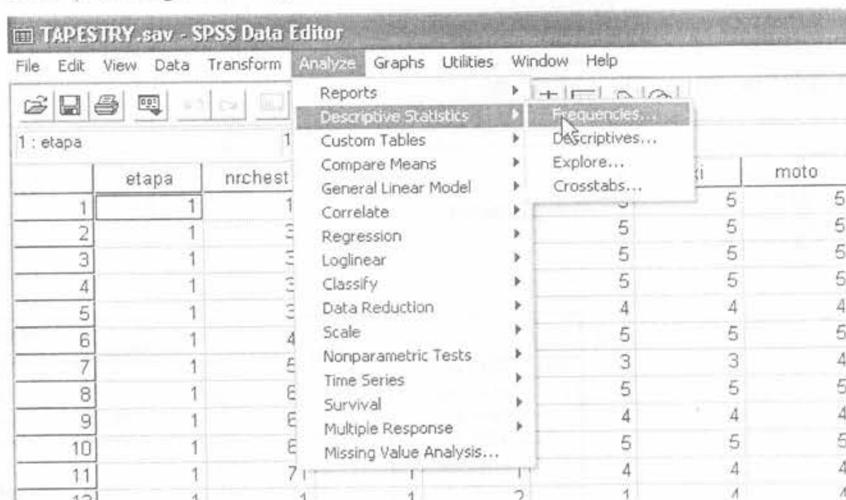


Figura 3.12 *Selectarea opţiunii* Analyze → Descriptive Statistics → Frequencies

Lista sursă a variabilelor este afişată în caseta din stânga a ferestrei *Frequencies.* Din această listă, se selectează variabila dorită, prin clic de

mouse, când bara de selecţie este poziţionată pe numele acesteia. Variabila selectată este mutată, prin clic pe butonul săgeată, din lista sursă în caseta *Variabile(s)*.

Observaţie! Acelaşi efect se obţine şi prin dublu clic de mouse de pe numele variabilei dorite.

Apoi, prin butonul de comandă *OK* se obţine *Tabelul de frecvenţă*, afişat în fereastra de rezultate *Output Viewer* (vezi figura 3.14).

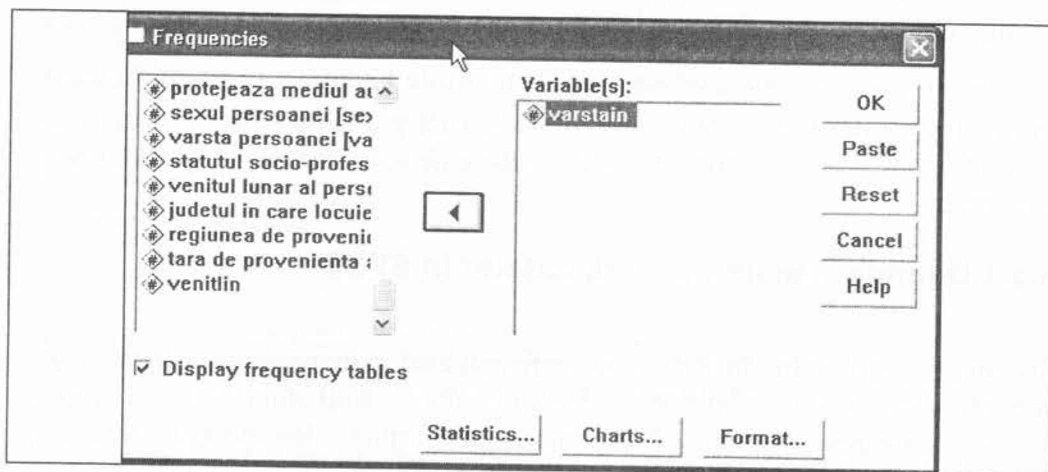


Figura 3.13 *Fereastra* Frequencies

Output
Frequencies
Title
Notes
VARSTAIN

Frequencies

VARSTA PELERINILOR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16-17 ani	24	6,0	6,0	6,0
	18-24 ani	88	22,0	22,0	28,0
	25-34 ani	81	20,3	20,3	48,3
	35-44 ani	75	18,8	18,8	67,0
	45-55 ani	70	17,5	17,5	84,5
	55-64	40	10,0	10,0	94,5
	65 ani si peste	22	5,5	5,5	100,0
	Total	400	100,0	100,0	

Figura 3.14 *Output-ul* Frequencies

3.3.2 Tabelul de frecvenţă

Într-un tabel de frecvenţă sunt prezentate, pentru fiecare variabilă selectată, următoarele elemente:

- *valorile sau clasele de valori ale variabilei, efectivul*;
- *procentele*;
- *procentele cumulate* corespunzătoare.

Pentru exemplificare, considerăm baza de date *tapestry.sav*-Iaşi, din care a fost selectată variabila numerică *Vârsta*. În urma aplicării demersului sistematizării datelor, descris în paragraful anterior, a rezultat *output*-ul prezentat în Figura 3.14.

Observaţie! Elementele tabelului (titlul, sursa etc.) pot fi completate, modificate, în funcţie de opţiuni, folosind fereastra de rezultate (*Output Viewer*). În acest scop, prin dublu clic selectăm tabelul pe care dorim să-l completăm şi efectuăm asupra sa sau asupra unui element din tabel operaţia necesară. Tabelul completat cu elementele sale se prezintă aşa cum se poate vedea în tabelul 3.1.

Tabelul 3.1 *Distribuţia după vârstă a eşantionului de pelerini la Sfânta Cuvioasă Parascheva, Iaşi, Octombrie 2002*

Vârsta	Frequency	Percent	Valid Percent	Cumulative Percent
16-17 ani	24	6.0	6.0	6.0
18-24 ani	88	22.0	22.0	28.0
25-34 ani	81	20.3	20.3	48.3
35-44 ani	75	18.8	18.8	67.0
45-55 ani	70	17.5	17.5	84.5
55-64 ani	40	10.0	10.0	94.5
≥65 ani	22	5.5	5.5	100.0
Total	400	100.0	100.0	

Sursa : *Calculat cu SPSS pe baza datelor TAPESTRY-Iaşi, oct. 2002*

3.3.3 Tabelul de contingență

Tabelul de contingență se obține în cazul unei variabile nominale (categoriale), procedându-se în mod asemănător cu tabelul de frecvență. Tabelul de contingență prezintă efectivul, procentele și procentele cumulate corespunzătoare fiecărei categorii a variabilei nominale (vezi tabelul 3.2).

Tabelul 3.2 *Distribuția pe sexe a eșantionului de pelerini la Sfânta Cuvioasă Parascheva, Iași, Octombrie 2002*

Sexul persoanei	Frequency	Percent	Valid Percent	Cumulative Percent
masculin	170	42.5	42.5	42.5
feminin	230	57.5	57.5	100.0
Total	400	100.0	100.0	

Sursa : *Calculat cu SPSS pe baza datelor TAPESTRY-Iași, Oct. 2002*

3.3.4 Tabelul de asociere (*Crosstabs*)

Acest tip de tabel este folosit pentru prezentarea relațiilor dintre două variabile categoriale. În fiecare rubrică (celulă) a tabelului, este prezentată frecvența parțială (n_{ij}), adică efectivul care poartă simultan o valoare a fiecărei variabile.

Obținerea unui tabel de asociere în SPSS presupune alegerea opțiunii *Crosstabs*, subordonată comenzii *Descriptive Statistics*, din meniul *Analyze* (vezi figura 3.15).

După ce se selectează această opțiune, apare pe monitor fereastra *Crosstabs*. În partea stângă a acestei ferestre se găsește lista sursă (lista tuturor variabilelor din baza de date), din care selectăm variabile pentru rânduri și variabile pentru coloane (vezi figura 3.16). În exemplul dat, s-au considerat variabilele: „sexul persoanei" și „categoria de vârstă".

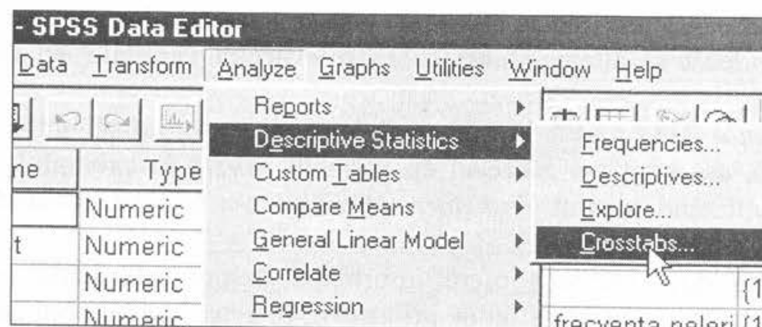
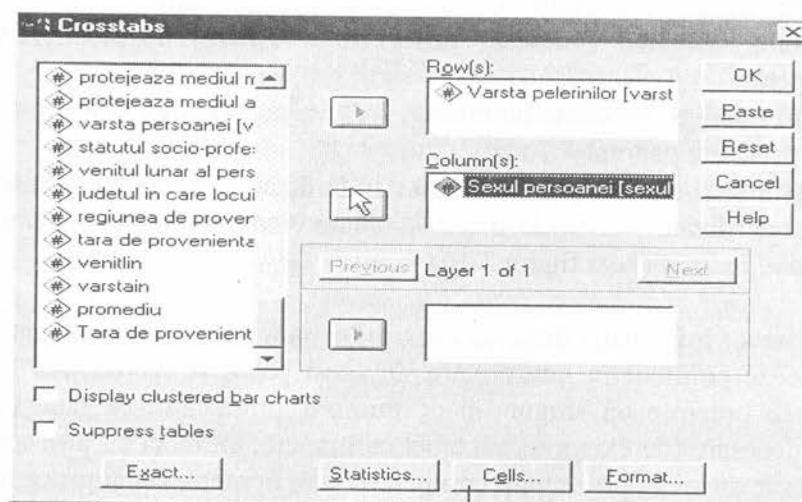


Figura 3.15 *Alegerea opțiunii* Analyze → Descriptive Statistics → Crosstabs



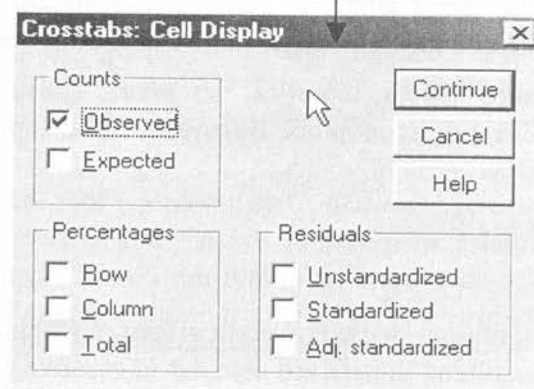


Figura 3.16 *Fereastra* Crosstabs

Observaţie 1. Dacă într-un *crosstabs* numărul categoriilor unei variabile este mai mare decât al alteia, atunci categoriile acelei variabile se plasează pe rânduri.

Observaţie 2. O variabilă numerică poate fi transformată într-o variabilă categorială, aşa cum s-a procedat cu variabila *vârstă* în exemplul considerat anterior, utilizând meniul *Transform*, din care s-a selectat comanda *Recode* (vezi paragrafele 3.4 şi 5.4.1).

O celulă din *crosstabs* oferă informaţia despre intersecţia celor două variabile. Pentru a obţine astfel de informaţii, se activează butonul de comandă *Cell* din fereastra *Crosstabs*, care are ca efect afişarea ferestrei *Crosstabs*: *Cell Display*. Astfel, se poate selecta forma sub care dorim să obţinem informaţia din fiecare celulă din crosstabs. Opţiunile sunt grupate astfel:

- *Count*: efective (numere) observate – *Observed*, efective teoretice (sperate) – *Expected*;
- *Percentages*: procente pe rânduri – *Row*, procente pe coloane – *Column* şi procente pe total – *Total*;
- *Residuals*: abateri (reziduuri) nestandardizate – *Unstandardized*, abateri standardizate – *Standardized* şi abateri standardizate ajustate – *Adj. standardized* (vezi figura 3.16).

Observaţie! Informaţia dintr-un *crosstabs* trebuie să fie uşor de analizat şi să nu se creeze confuzii în interpretare. În acest scop, se recomandă ca într-un tabel să se prezinte un singur tip de numere, producând pe rând (succesiv) tabelele necesare (de exemplu, un tabel cu numere, alt tabel cu procente). Dacă este necesar totuşi să prezentăm comparativ atât numere, cât şi procente, atunci acestea se trec unele lângă altele, în ordinea importanţei lor în interpretare. Vom exemplifica diferite tabele posibile pentru cele două variabile selectate.

Observaţie! Pentru partea teoretică, se poate consulta Elisabeta Jaba, *Statistica*, ediţia a 3-a, Ed. Economică, Bucureşti, 2002, capitolul 3).

3.3.5 Exemple

1. În tabelul 3.3, eşantionul de pelerini este grupat după clasa de vârstă şi după sexul persoanei. Rezultatul clasificării este dat în efective (numeric). Astfel, în eşantionul observat, sunt 47 de pelerini de sex masculin în vârstă de până la 25 de ani.

Tabelul 3.3 Vârsta pelerinilor * Sexul persoanei Crosstabulation. Count

Vârsta pelerinilor	Sexul persoanei		Total
	masculin	feminin	
< 25 ani	47	65	112
25-64 ani	113	153	266
≥65 ani	10	12	22
Total	170	230	400

2. În tabelul 3.4 este prezentat un *crosstabs* cu rezultatul procentual pe rânduri. Acest rezultat se obține divizând fiecare număr prin totalul rândului căruia îi aparține. De exemplu, 42% din totalul pelerinilor tineri sunt de sex masculin, iar 58% sunt de sex feminin. Rezultatul poate fi interpretat ca *probabilitate condiționată* exprimând, de exemplu, probabilitatea pelerinilor de sex masculin de a se găsi în grupa tânără.

Tabelul 3.4 Vârsta pelerinilor * Sexul persoanei Crosstabulation % within Vârsta pelerinilor

Vârsta pelerinilor	Sexul persoanei		Total
	masculin	feminin	
< 25 ani	42.0%	58.0%	100.0%
25-64 ani	42.5%	57.5%	100.0%
65 și peste	45.5%	54.5%	100.0%
Total	42.5%	57.5%	100.0%

3. În tabelul 3.5 este prezentat un *crosstabs* cu rezultatul procentual pe coloane. Se obține divizând fiecare număr prin totalul coloanei căreia îi aparține. De exemplu, 27,6% din totalul pelerinilor de sex masculin fac parte din grupa tânără. Rezultatul poate fi interpretat ca o *probabilitate condiționată*, exprimând, de exemplu, șansa pelerinilor din grupa tânără de a aparține grupei masculine.

Observație! Se face distincție între cele două tipuri de probabilități condiționate. Astfel, numai 27,6% din pelerinii bărbați aparțin grupei tinere, dar în grupa pelerinilor tineri bărbații reprezintă 42% .

Tabelul 3.5 Vârsta pelerinilor * Sexul persoanei Crosstabulation % within Sexul persoanei

Vârsta pelerinilor	Sexul persoanei		Total
	masculin	feminin	
< 25 ani	27.6%	28.3%	28.0%
25-64 ani	66.5%	66.5%	66.5%
65 şi peste	5.9%	5.2%	5.5%
Total	100.0%	100.0%	100.0%

4. În tabelul 3.6 sunt prezentate *frecvenţele relative procentuale parţiale*. Se obţin prin divizarea fiecărui număr din totalul eşantionului. Rezultatul poate fi interpretat ca probabilitatea fiecărui pelerin de a răspunde la ambele condiţii, de a aparţine la o grupă de vârstă şi de a fi de un anumit sex. De exemplu, 11,8% din totalul eşantionului de pelerini au şansa de a fi tineri şi de sex masculin.

Tabelul 3.6 Vârsta pelerinilor * Sexul persoanei Crosstabulation% of Total

Vârsta pelerinilor	Sexul persoanei		Total
	masculin	feminin	
< 25 ani	11.8%	16.3%	28.0%
25-64 ani	28.3%	38.3%	66.5%
65 şi peste	2.5%	3.0%	5.5%
Total	42.5%	57.5%	100.0%

5. În tabelul 3.7 este prezentat un *crosstabs* cu rezultatul într-o formă combinată, numere şi procente (clasificarea pelerinilor simultan după cele două variabile faţă de totalul rândului căruia îi aparţin).

Tabelul 3.7 Vârsta pelerinilor * Sexul persoanei Crosstabulation

Vârsta pelerinilor		Sexul persoanei		Total
		masculin	feminin	
< 25 ani	Count	47	65	112
	%within Vârsta pelerinilor	42.0%	58.0%	100.0%

25-64 ani	Count	113	153	266
	%within Vârsta pelerinilor	42.5%	57.5%	100.0%
65 și peste	Count	10	12	22
	%within Vârsta pelerinilor	45.5%	54.5%	100.0%
Total	Count	170	230	400
	%within Vârsta pelerinilor	42.5%	57.5%	100.0%

3.4 Transformarea datelor

3.4.1 Recodificarea variabilelor folosind comanda RECODE

Recodificarea variabilelor este o modalitate de transformare a unei variabile prin combinarea valorilor acesteia într-un număr mai mic de categorii. De exemplu, vârsta exprimată în ani pentru fiecare subiect poate fi regrupată pe categorii: „tânăr", „adult", „în vârstă". Pentru a realiza această transformare în SPSS, este necesar:

1. să se decidă numărul de grupe (de regulă, se limitează la 3 sau 4 grupe);
2. să se verifice dacă fiecare din vechile valori se poate combina în noile valori.

Pentru exemplificare, folosim datele din tabelul 3.8.

Tabelul 3.8 *Transformarea datelor prin recodificare*

Vârsta respondentului codificată inițial în foaia de date	Noua categorie
18-20 ani	Tânăr (1)
21-60 ani	Adult (2)
61 ani și peste	În vârstă (3)

Pentru început, atribuim un nume nou unei variabile. De exemplu, *Vârsta* o recodăm în *Vârsta1*.

În continuare, se activează comanda *Recode* din meniul *Transform*. Această comandă are în subordine două opţiuni (vezi figura 3.18):

- *Recode Into Different Variables* (pentru recodificarea într-o variabilă diferită);
- *Recode Into Same Variables* (pentru recodificarea în aceeaşi variabilă).

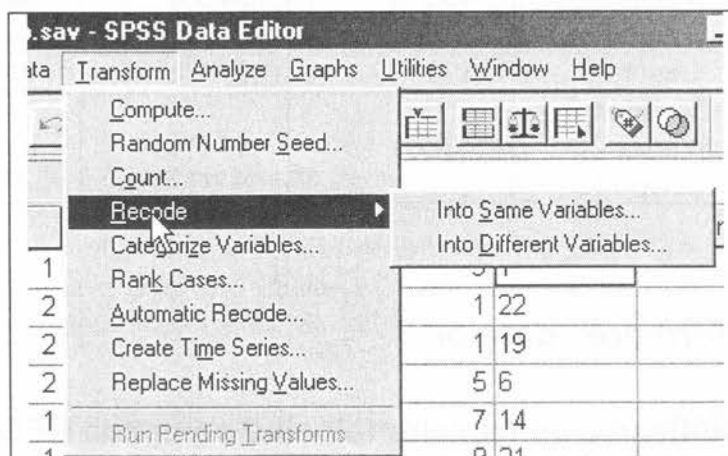


Figura 3.18 *Comanda* Transform → Recode

Recodificarea într-o variabilă diferită

Pentru o astfel de recodificare, se selectează opţiunea *Into Different Variables*, care deschide fereastra *Recode into Different Variables* (vezi figura 3.19).

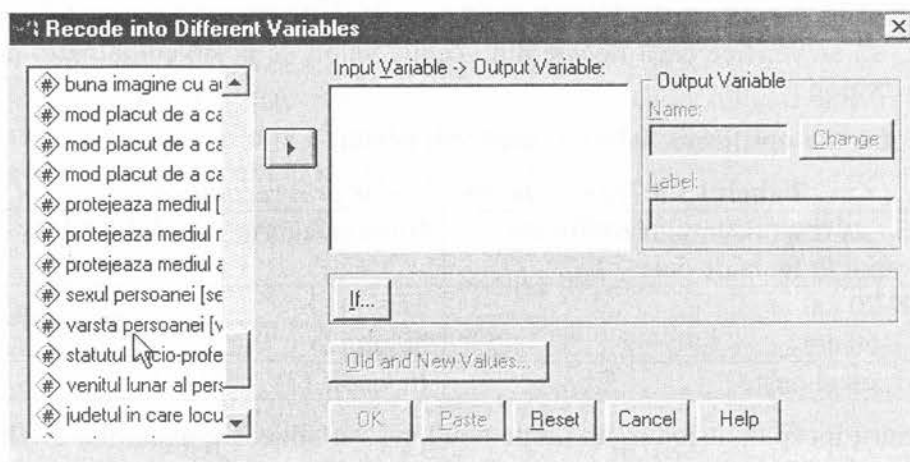


Figura 3.19 *Fereastra* Recode into Different Variables

În fereastra *Recode into Different Variables* se parcurg următorii pași (vezi figura 3.20):

- se selectează variabila pe care dorim să o recodificăm, de exemplu, *Vârsta*, în lista variabilelor și se mută în lista variabilelor de recodat;
- se scrie numele noii variabile, *Vârsta1*, în caseta *Name* din zona *Output Variabile*;
- se scrie în caseta *Label* eticheta noii variabile;
- se activează butonul de comandă *Change* pentru a fi operată modificarea numelui variabilei.

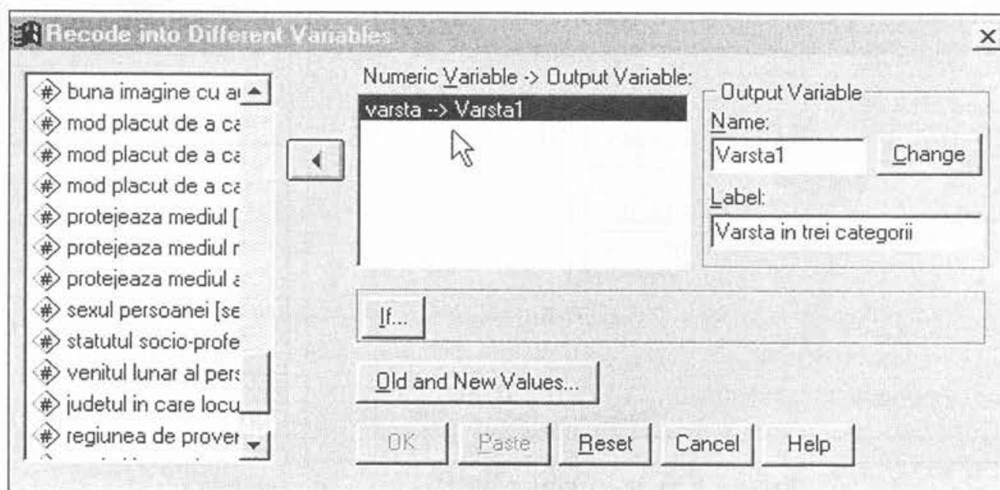


Figura 3.20 *Schimbarea numelui variabilei* Vârstă *în* Vârstă1

Tot în fereastra *Recode into Different Variables*, se definesc categoriile pentru variabila numerică. Pentru aceasta, se acționează butonul de comandă *Old and New Values* care deschide fereastra *Recode into Different Variables: Old and New Values* (vezi figura 3.21).

În funcție de opțiunea dorită, efectuăm un set de operații.

Pentru a *schimba o valoare particulară într-o valoare nouă*, se introduce valoarea veche în caseta *Old Value* și valoarea nouă în caseta *New Value* și apoi se acționează butonul de comandă *Add*. De regulă, se schimbă o valoare reală cu altă valoare reală. De exemplu, se schimbă 21-60 ani în valoarea *2*, adică se combină toate vârstele de 21-60 ani într-o singură valoare, *2*.

În acest scop, selectăm butonul de opțiuni *Range*. Casetele de editare sunt folosite pentru a stabili limita inferioară și respectiv limita superioară a intervalului dorit (*through* – de la-până la). Se scrie 21 în caseta din stânga și 60 în cea din dreapta. Apoi se selectează butonul de opțiuni *Value* din zona

New Value şi se scrie *2* în caseta de editare, după care se acţionează butonul *Add*. Se procedează în mod analog pentru toate categoriile (vezi figura 3.21).

Prin clic pe butonul de comandă *Continue*, se revine în fereastra *Recode into Different Variables*. Prin butonul de comanda *OK*, se va declanşa recodificarea variabilei. Noua variabilă apare în foaia de date, *Date View* (vezi figura 3.21), cu datele de cod corespunzătoare fiecărui caz.

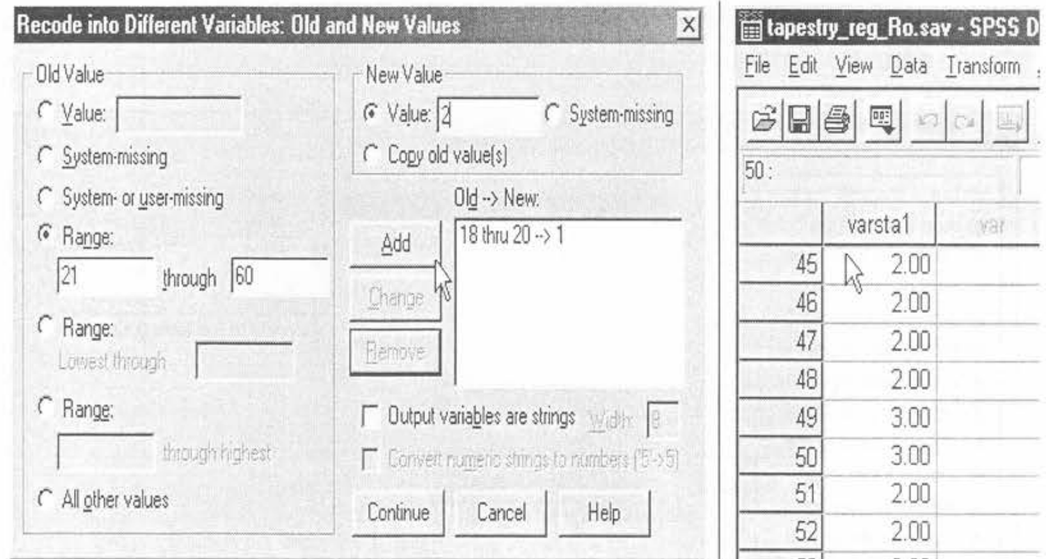


Figura 3.21 *Recodificarea valorilor unei variabile*

3.4.2 Crearea unei noi variabile folosind comanda COMPUTE

Se pot crea noi variabile plecând de la variabilele vechi, folosind din meniul *Transform* comanda *Compute*.

Din fişierul *dez_reg.sav*, vom considera variabile PIB_98, PIB_99, PIB_00. Pe baza lor, se pot crea noi variabile, de exemplu, indicatori utilizaţi în analiza seriilor de timp (indicele de variaţie cu bază mobilă, indicele de variaţie cu bază fixă, sporul cu bază mobilă, sporul cu bază fixă etc.).

Comanda *Compute* din meniul *Transform* deschide fereastra *Compute Variable* (vezi figura 3.22).

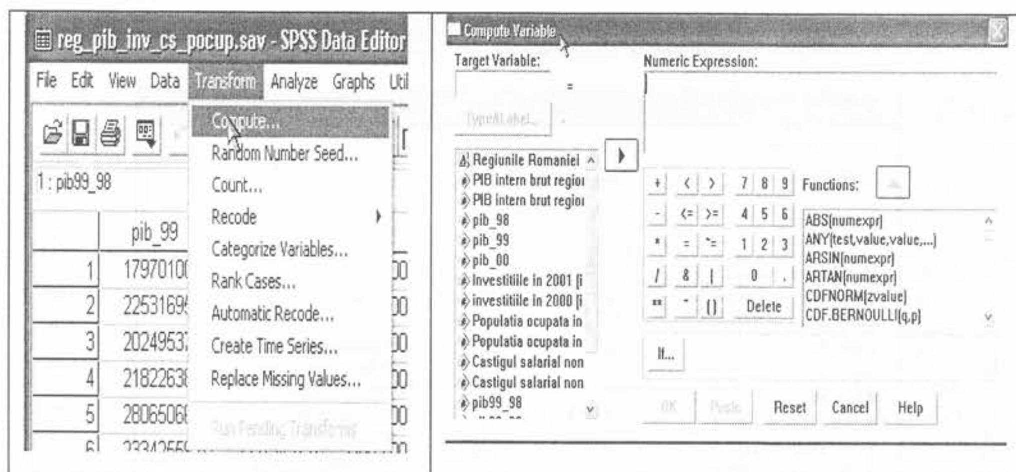


Figura 3.22 *Deschiderea ferestrei* Compute Variable

În fereastra *Compute Variable*, se parcurg următorii paşi:

- în caseta *Target Variable*, se scrie numele noii variabile, de exemplu PIB99_98;
- în caseta *Numeric Expression*, se introduce formula de calcul pentru această nouă variabilă. Această operaţie se poate efectua pe două căi:
  - se selectează prima variabilă din lista variabilelor şi se mută, cu ajutorul butonului săgeată, în caseta *Numeric Expression*; operatorii, operanzii şi eventual funcţiile folosite în formulă se selectează prin clic de mouse de pe butoanele corespunzătoare. De exemplu, pentru împărţire, se selectează semnul „/", care se mută în caseta expresiei numerice (vezi figura 3.23). Prin butonul *OK*, comanda este preluată de SPSS, iar noua variabilă se poate vedea în fereastra *editorul de date*;
  - se introduce de la tastatură formula direct în caseta *Numeric Expression*.

Pentru noua variabilă, rezultată din raportarea celor două variabile, se poate introduce, în foaia *Variable View*, numele complet, de exemplu, Indicele PIB 1999/1998.

Se poate, de asemenea, crea o nouă variabilă, modificând o variabilă veche prin multiplicarea sau reducerea valorilor acesteia cu o constantă etc.

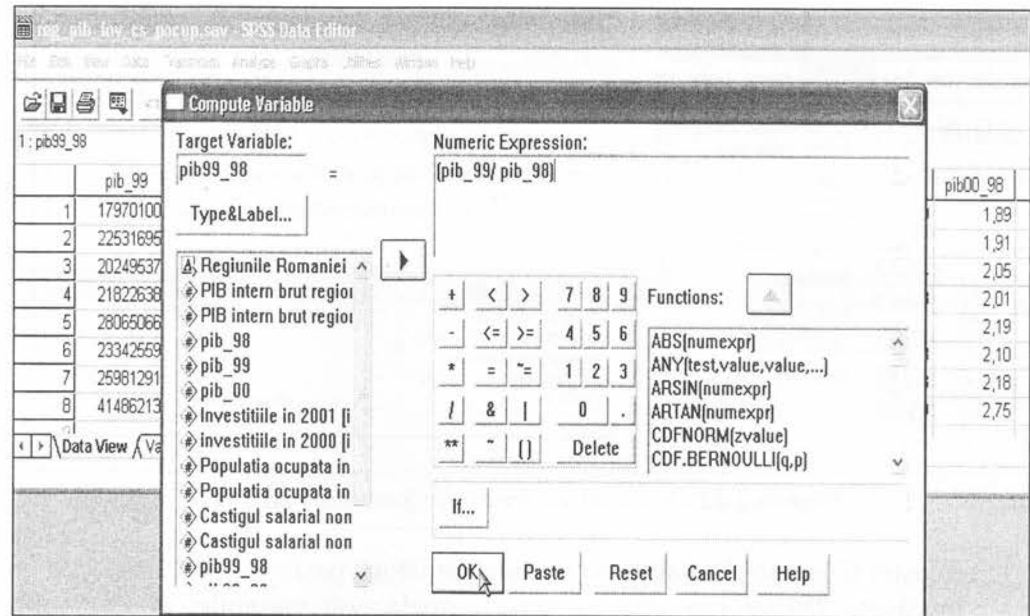


Figura 3.23 *Calculul unei noi variabile*, PIB99_98, *prin* Compute Variable

3.5 Modificarea unui tabel în SPSS

În paragraful 3.3.2 s-a menţionat cum se poate *completa un tabel cu elementele* necesare interpretării corecte a distribuţiei prezentate. În continuare, vom preciza cum se poate *modifica un tabel în SPSS*. O astfel de operaţie presupune parcurgerea următorilor paşi:

- dublu clic asupra tabelului afişat în fereastra de rezultate *Output – SPSS Viewer*. În felul acesta se selectează tabelul pe care dorim să-l modificăm şi, totodată, se afişează bara cu instrumentele *Formatting*, iar în bara meniu apare meniul *Pivot* (vezi figura 3.24);

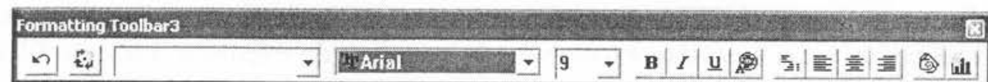


Figura 3.24 *Bara de instrumente* Formatting

- se selectează din meniul *Pivot* comanda *Pivoting Trays* sau din bara de instrumente *Formatting*, pictograma *Pivotting*, care deschide fereastra *Pivoting Trays* (vezi figura 3.25);

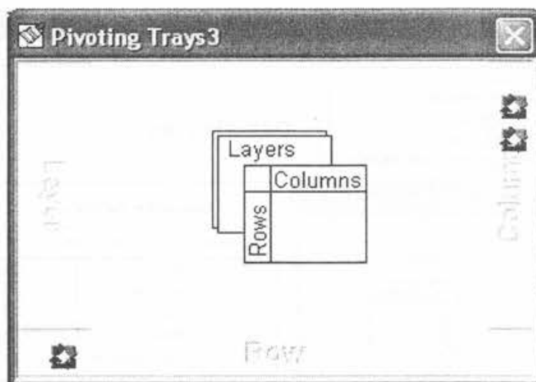


Figura 3.25 *Fereastra* Pivoting Trays

- se schimbă locul pictogramelor floare, prin „tragerea" lor (tehnica *drag&drop*) de pe rânduri pe coloane și invers, în funcție de ce variabile dorim să fie schimbate (vezi figura 3.26). Ca urmare a acestei operații, se produce modificarea tabelului selectat (vezi figura 3.27).

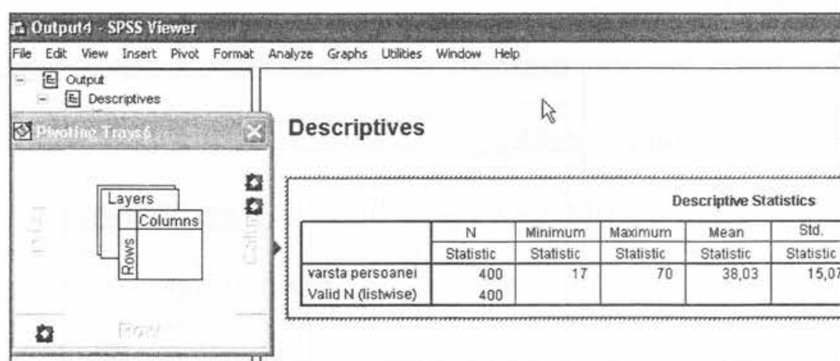


Figura 3.26 *Caseta* Pivoting Trays *de modificare a tabelului de rezultate*

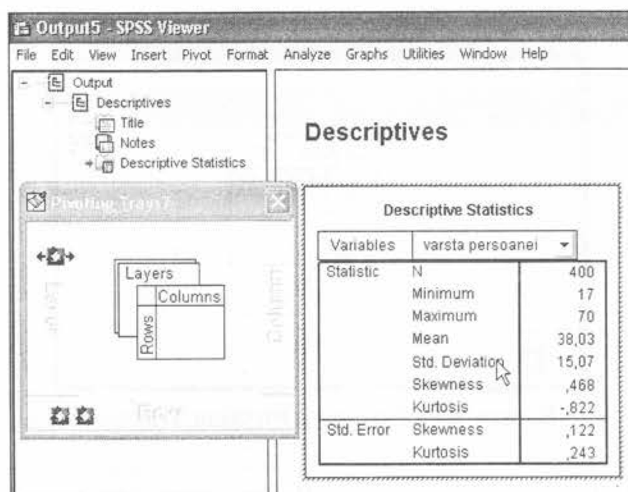


Figura 3.27 *Caseta* Pivoting Trays, *cu icoanele schimbate și tabelul modificat*

O altă modalitate de modificare a unui tabel o reprezintă folosirea meniului rapid apelat din fereastra de rezultate, când mouse-ul este plasat pe tabelul dorit (vezi figura 3.28).

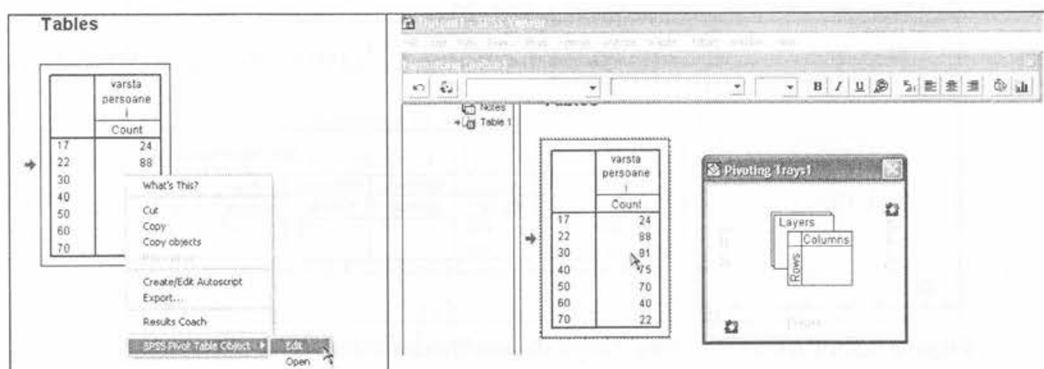


Figura 3.28 *Apelarea meniului rapid pentru modificarea unui tabel*

CAPITOLUL 4

REPREZENTAREA GRAFICĂ A UNEI DISTRIBUȚII ÎN SPSS

- **Elemente introductive**
- **Grafice pentru distribuții după o variabilă cantitativă**
- **Grafice pentru distribuții după o variabilă calitativă (nominală)**
- **Grafice pentru distribuții bivariate**
- **Modificarea unui grafic în SPSS**

4.1 Elemente introductive

Graficele prezintă în mod sintetic, sub formă vizuală, o distribuție statistică. Cu ajutorul graficelor se poate avea, dintr-o singură privire, o viziune de ansamblu asupra datelor.

4.1.1 Elementele unui grafic

Pentru a fi uşor de interpretat, un grafic, pe lângă diagrama propriu-zisă, trebuie să aibă precizate următoarele elemente:

- *titlul graficului* oferă informații asupra fenomenului reprezentat. Titlul graficului coincide cu titlul tabelului de date;
- *axele de coordonate* sunt folosite pentru a reprezenta variabilele (cu unitățile de măsură corespunzătoare) şi au scala de măsură precizată. Pe abscisă se înscrie variabila de distribuție (variabila independentă), iar pe ordonată frecvența (sau variabila dependentă);
- *legenda* este folosită pentru a explica elementele din diagramă;
- *sursa* precizează originea datelor reprezentate.

Citirea unui grafic presupune observarea şi interpretarea diagramei atât în ansamblul ei, cât şi a oricărui punct de pe diagramă. Fiecare punct de pe grafic reprezintă relația dintre variabilele considerate pe axe, ceea ce implică verificarea gradației axelor (dacă încep sau nu cu zero).

Observație! Pentru mai multe detalii, vezi Elisabeta Jaba, *Statistica*, ediția a III-a, Editura Economică, Bucureşti, 2002, pp. 59-60.

4.1.2 Tipuri de grafice

Alegerea graficului pentru reprezentarea unei distribuții se face în funcție de scopul urmărit şi depinde esențial de numărul variabilelor considerate, precum şi de tipul acestora. *SPSS* oferă o paletă foarte largă de tipuri de grafice. Acestea, precum şi modalitățile de obținere a lor sunt organizate, în principal, în meniul *Graphs* (vezi figura 4.1).

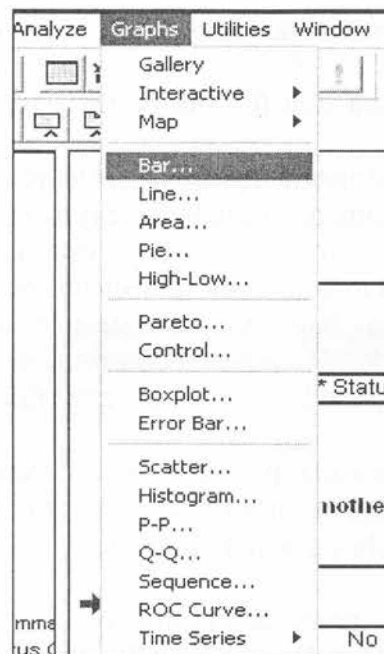


Figura 4.1 *Meniul* Graphs

De asemenea, pot fi obţinute grafice şi cu ajutorul butoanelor de comandă *Charts* sau *Plots*, prezente în anumite ferestre de dialog, deschise de comenzile meniului *Analyze*. De exemplu, fereastra *Frequencies: Charts* (vezi figura 4.2).

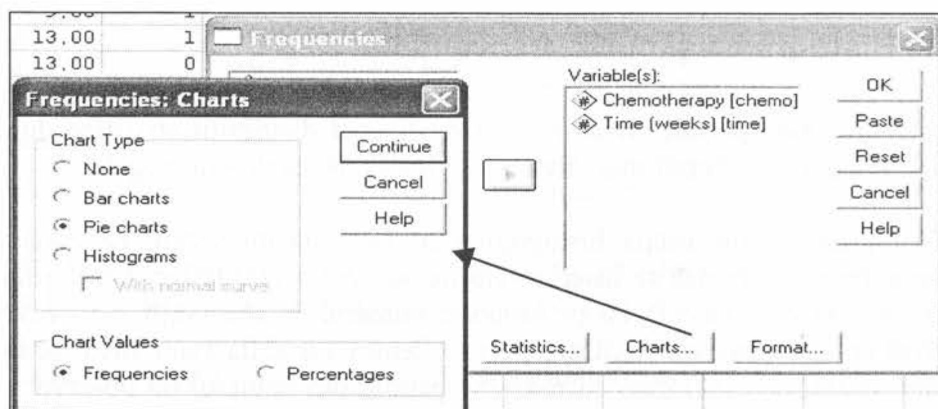


Figura 4.2 *Fereastra* Frequencies: Charts

Cele mai uzuale grafice din meniul *Graphs* sunt cele prezentate mai jos.

Bar. Diagrama în bare este folosită pentru a reprezenta grafic mediile diferitelor grupe dintr-o colectivitate (*Summaries for groups of cases*) sau

valorile medii ale diferitelor variabile pentru aceeaşi colectivitate (*Summaries of separate variables*).

Line. Diagrama liniară este folosită pentru a reprezenta, de regulă, valori medii.

Pie. Diagrama de structură „plăcintă" este folosită pentru reprezentarea frecvenţelor absolute (numere) şi/sau relative (procente) pe categorii/grupe.

Boxplot. Diagrama „cutia cu mustăţi" este folosită pentru a prezenta amplitudinea, intervalul interquartilic şi mediana unei distribuţii.

Error Bar. Diagrama „bara erorilor" este folosită pentru a arăta media şi intervalul de încredere de 95% pentru media respectivă.

Scatter. Diagrama „norul de puncte" este folosită pentru a reprezenta relaţiile între variabile.

Histograma. Este folosită pentru a arăta forma unei distribuţii după o variabilă înregistrată asupra unei colectivităţi (frecvenţele de apariţie pentru diferite clase de valori ale variabilei observate).

În continuare, vom prezenta construcţia câtorva diagrame folosind fie meniul *Graphs*, fie butoanele de comandă *Charts* sau *Plots* din ferestrele de dialog subordonate anumitor comenzi din meniul *Analyze*.

4.2 Grafice pentru distribuţii după o variabilă cantitativă

4.2.1 Histograma şi curba frecvenţelor

Aceste diagrame permit vizualizarea formei unei distribuţii statistice după o *variabilă cantitativă continuă*, divizată pe intervale, egale sau inegale.

Histograma. Construcţia histogramei se face într-un sistem de două axe rectangulare: pe abscisă se înscriu valorile variabilei cantitative, sub formă de intervale (clase de valori), iar pe ordonată numărul de observaţii sau frecvenţa relativă corespunzătoare fiecărui interval. Pentru variabila cantitativă, se ia un număr de intervale (k) egal cu rădăcina pătrată din numărul de observaţii (n) sau $k=1+3.322\ lgn$. Se recomandă să fie utilizată histograma când $n \geq 50$.

SPSS oferă mai multe modalităţi de obţinere a unei histograme. Prezentăm în continuare câteva dintre acestea.

a. Comanda *Histogram*, din meniul *Graphs*.

În fereastra de dialog *Histogram* (vezi figura 4.3), selectăm variabila pentru care dorim să construim histograma, prin clic asupra ei, şi o trecem în caseta *Variable*. Se poate adăuga *curba frecvențelor*, prin bifare în caseta de validare corespunzătoare (*Display normal curve*), apoi, prin activarea butonului de comandă *Titles*, se poate adăuga titlul dorit.

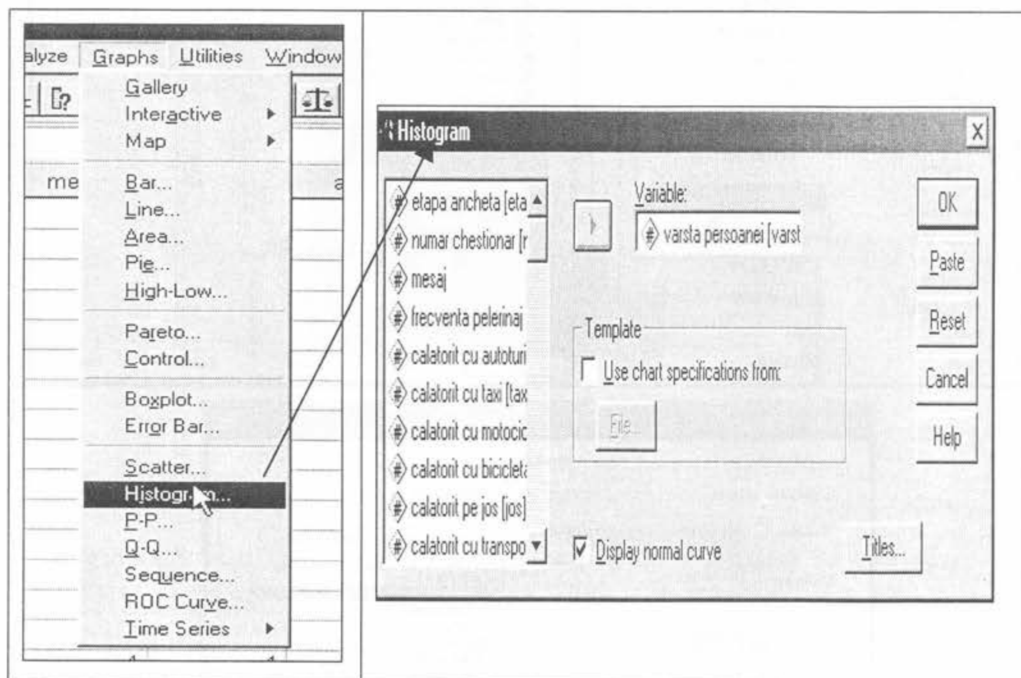


Figura 4.3 *Crearea histogramei prin: meniul* Graphs → *comanda* Histogram

Urmând acest algoritm, construim histograma pentru distribuția pelerinilor după vârstă, considerată în exemplul folosit anterior (vezi figura 4.5)

b. Comanda *Interactive* cu opțiunea *Histogram*, din meniul *Graphs*.

În fereastra de dialog *Create Histogram* (vezi figura 4.4), se alege variabila de reprezentat şi se mută, prin *tragere*, în caseta axei abscisă. Pe axa ordonatei se reprezintă numărul cazurilor în fiecare grupă (interval). Numărul cazurilor poate fi exprimat numeric sau procentual, folosind cadrul de pagină *Options*. După stabilirea opțiunilor, prin butonul de comandă *OK* se obține graficul în fereastra de rezultate *Output Viewer* (vezi figura 4.5).

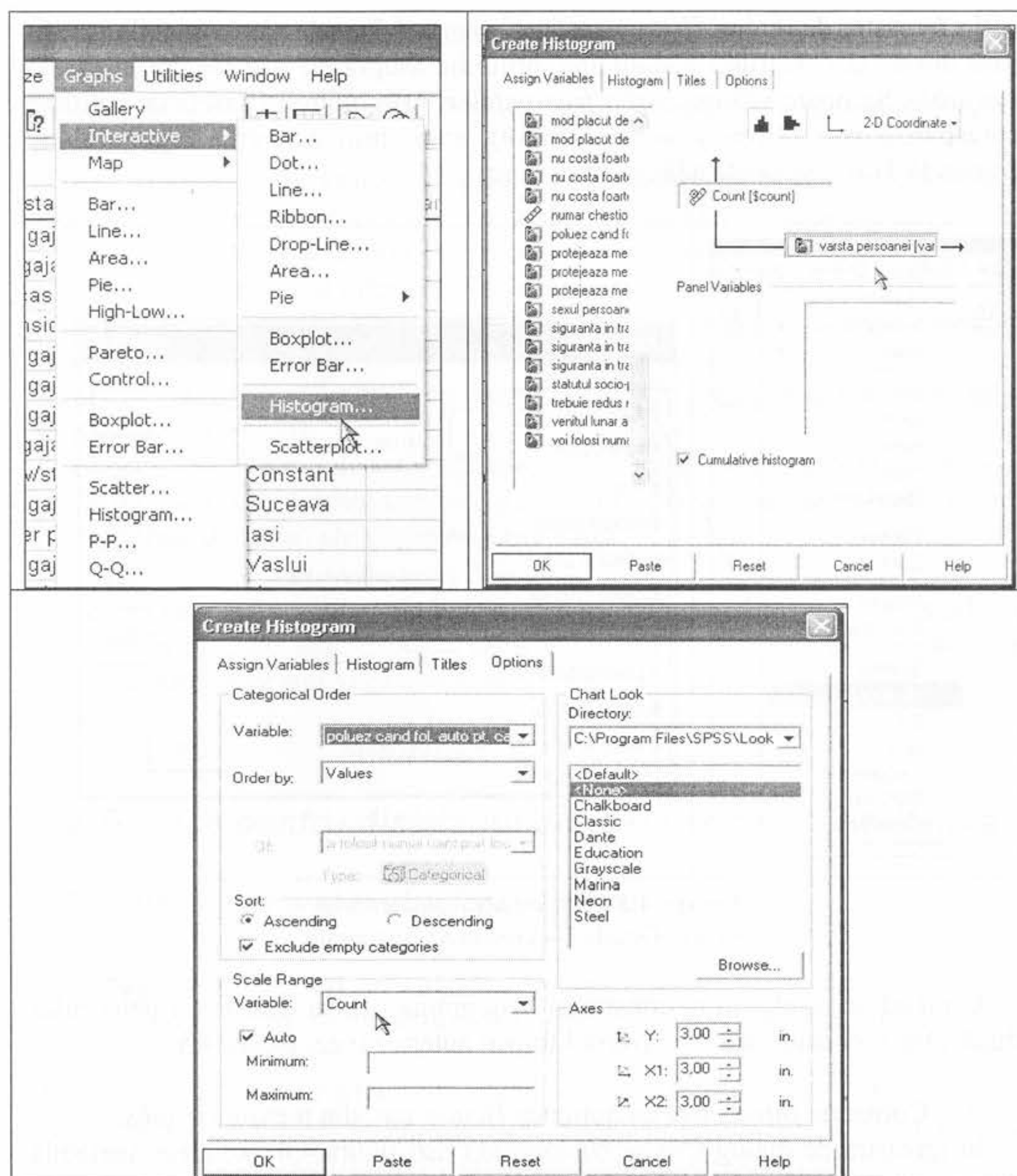


Figura 4.4 *Crearea histogramei prin:*
meniul Graphs → *comanda* Interactive → *opțiunea* Histogram

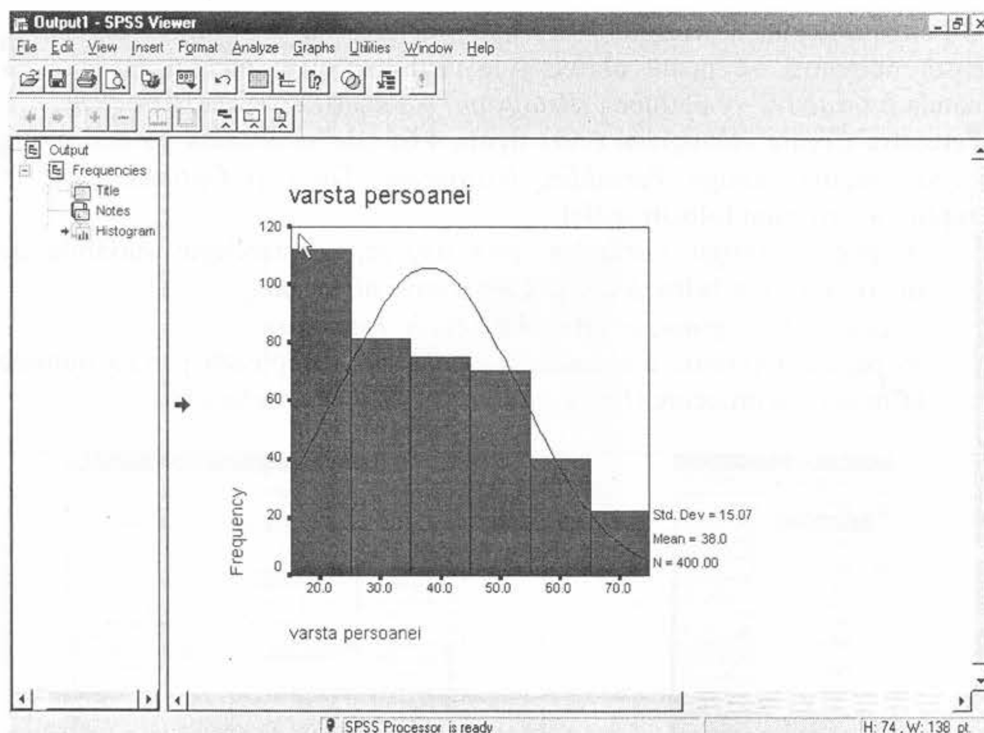


Figura 4.5 *Distribuția după vârstă a pelerinilor din eșantionul Tapestry-Iași, octombrie 2002*

Interpretare. Se observă că, pe ansamblu, eșantionul are o distribuție după vârstă asimetrică la dreapta, predominând vârsta tânără. Clasele 15-25 și 25-35 de ani au frecvențele cele mai mari, însumând împreună aproximativ jumătate din volumul eșantionului.

c. Comanda *Frequencies* din meniul *Graphs*.

O altă cale de construire a histogramei în SPSS este oferită de comanda *Frequencies*, prezentată în secțiunea precedentă, selectând succesiv: meniul *Analyze* ⟶ comanda *Descriptive Statistics* ⟶ opțiunea *Frequencies* ⟶ butonul de comandă *Charts* ⟶ butonul de opțiuni *Histogram.*

Curba frecvențelor. Curba frecvențelor se obține prin ajustarea histogramei și este folosită pentru verificarea normalității unei distribuții.

Curba frecvențelor poate fi suprapusă histogramei, reprezentând distribuția teoretică corespunzătoare, cu aceeași medie și aceeași varianță.

Un caz particular de curbă a frecvențelor este *curba frecvențelor cumulate*. Această diagramă se poate obține selectând succesiv: meniul *Graphs* → comanda *Interactive* → opțiunea *Histogram* → fereastra *Create Histogram*.

Fereastra *Create Histogram* (vezi figura 4.6) este structurată pe mai multe cadre de pagină: *Assign Variables*, *Histogram*, *Titles* și *Options*, care, în exemplul nostru, sunt folosite astfel:

- în pagina *Assign Variables*, prin tragere, se stabilește variabila de distribuție și se bifează caseta *Cumulative histogram*;
- în pagina *Histogram*, se bifează caseta *Normal curve*;
- în pagina *Options*, din zona *Scale Range*, se optează pentru numere (*Count*) sau procente (*Percent*) pentru axa Y (axa ordonatei).

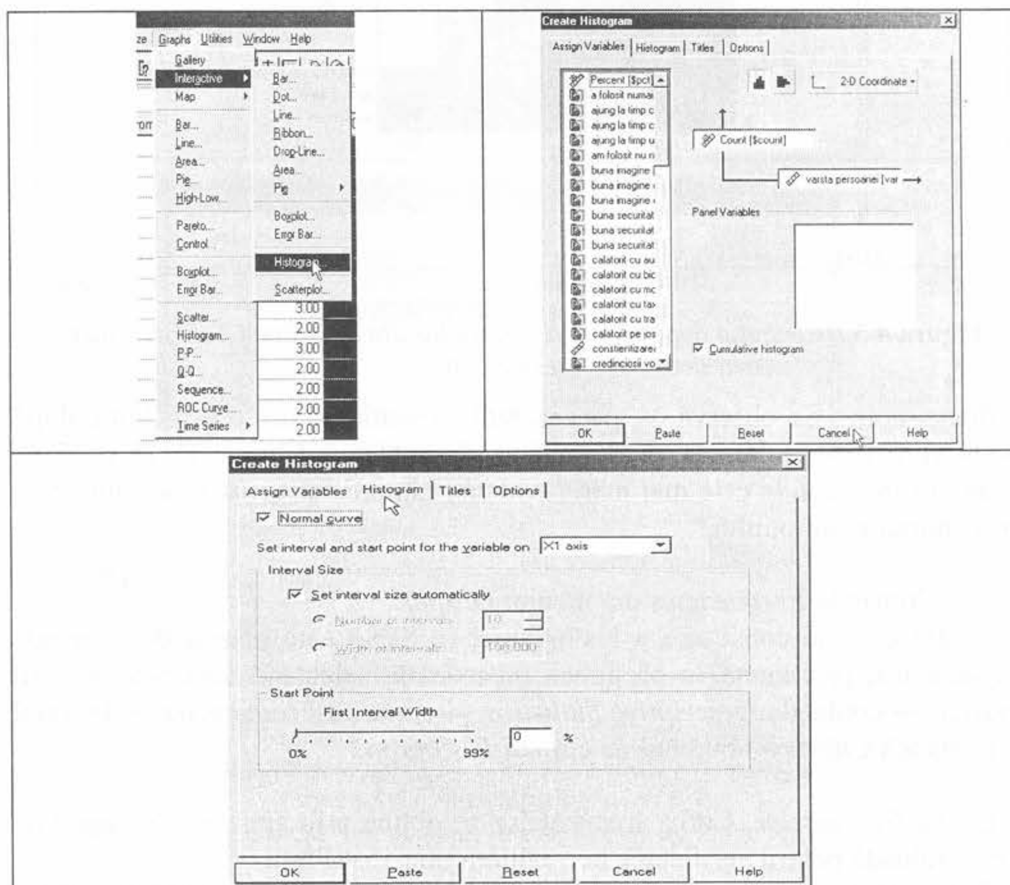


Figura 4.6 *Demersul folosit pentru construirea curbei cumulative*

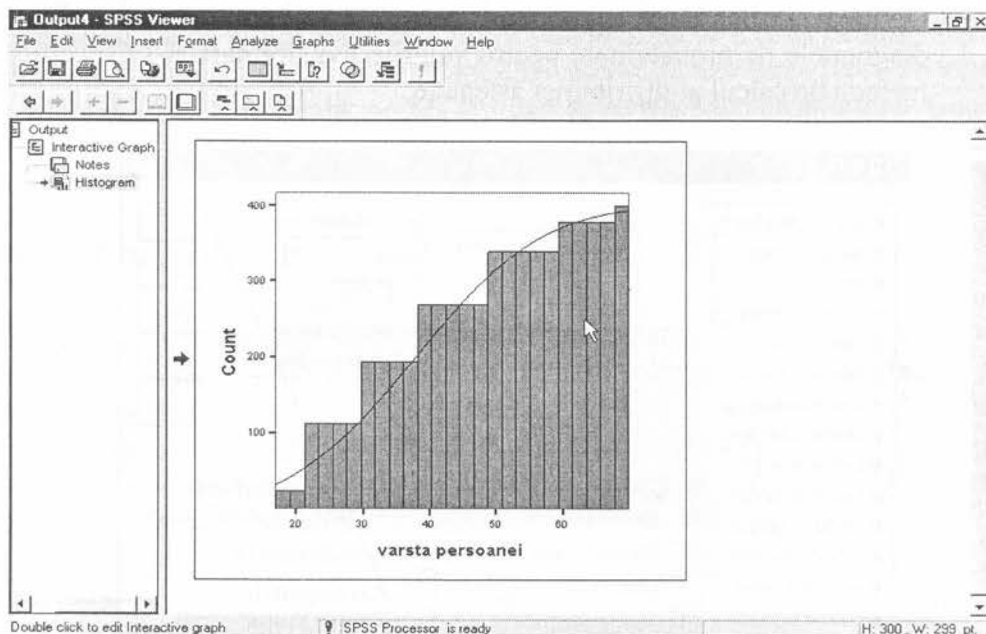


Figura 4.7 *Curba frecvențelor cumulate*

Interpretare. Frecvențele cumulate reprezintă efectivul care are o valoare mai mică decât limita superioară a intervalului (clasei) de variație. În histograma cumulată, fiecare bară reprezintă suma frecvențelor cumulate anterior acelui interval, plus frecvența intervalului curent. În curba frecvențelor cumulate, se citește pe ordonată, pentru fiecare punct de pe abscisă, o aproximare a frecvenței cumulate până la acel punct. De exemplu, curba cumulativă din figura 4.7 ne arată că până la 40 de ani sunt aproximativ 200 de persoane.

4.2.2 Q-Q Plot

Q-Q Plot este folosit pentru verificarea normalității. Demersul pentru construirea diagramei Q-Q plot presupune parcurgerea următorilor pași:

- Se alege din meniu *Graphs* comanda *Q-Q*, care deschide fereastra dialog *Q-Q plots* (vezi figura 4.8);
- Se alege o variabilă (sau mai multe variabile) și se mută în lista *Variabiles*;
- Se alege modelul distribuției test, în acest caz, *distribuția normală*;

- Opțional, se pot alege *căi de transformare* a variabilei pentru a obține diagramele de probabilitate pentru valorile transformate şi se specifică metoda de calcul al distribuției aşteptate.

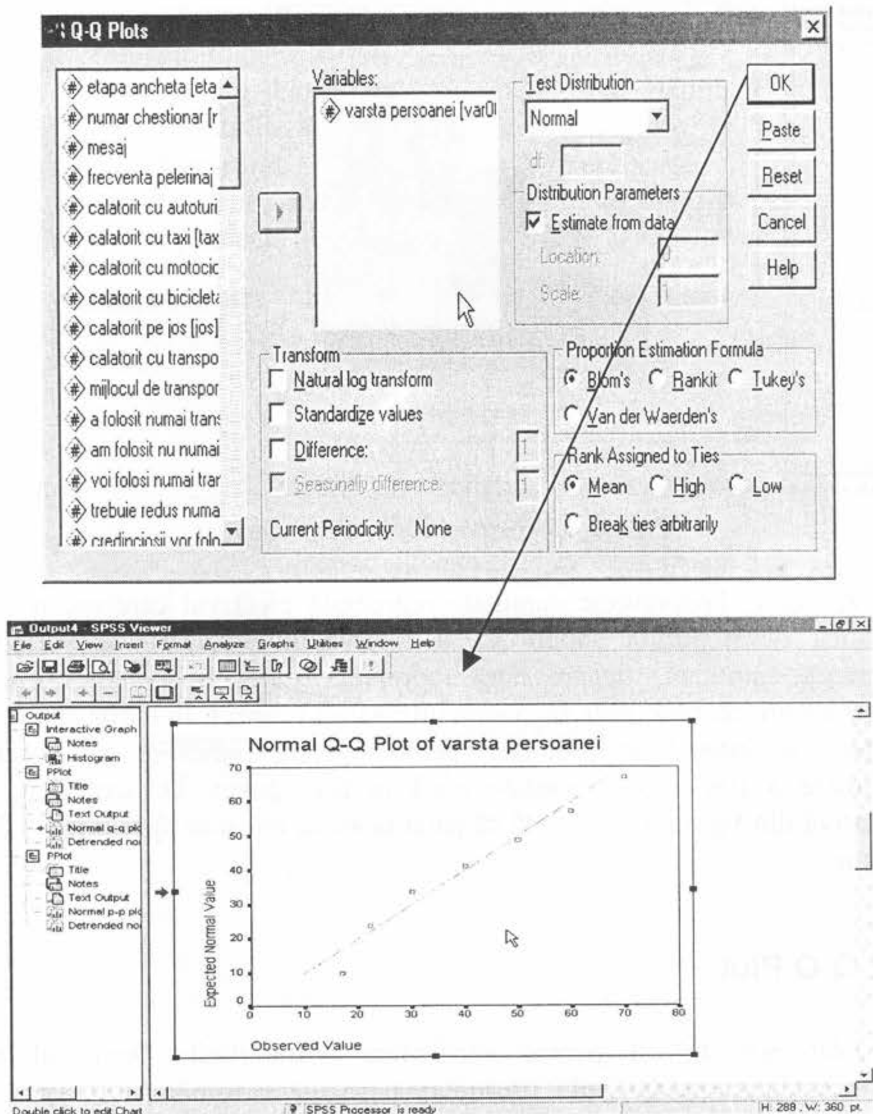


Figura 4.8 *Demersul folosit în construirea diagramei* Q-Q Plot *pentru verificarea normalității unei distribuții*

Un *Q-Q Plot* compară datele observate cu datele pe care ar trebui să le avem dacă distribuția noastră ar urma perfect o distribuție normală, cu aceeaşi medie

şi aceeaşi abatere standard. Valorile observate şi valorile sperate sunt comparate pe un grafic, unde pe abscisă sunt valorile observate pentru variabila X, iar pe ordonată sunt valorile variabilei Z corespunzătoare. Dacă distribuţia variabilei X ar fi normală, atunci graficul ar trebui să arate o tendinţă liniară (vezi figura 4.8).

4.2.3 Boxplot

Diagrama *Boxplot* este folosită pentru prezentarea unei distribuţii după o variabilă numerică, chiar atunci când numărul datelor de care dispunem este mic. Construcţia sa presupune ordonarea datelor şi împărţirea lor în patru grupe, fiecare grupă reprezentând 25% din distribuţie. Sunt marcate astfel cinci valori ale variabilei, şi anume: valoarea minimă şi valoarea maximă, fără outlieri, quartila 1, quartila 3 şi mediana (vezi figura 4.9) .

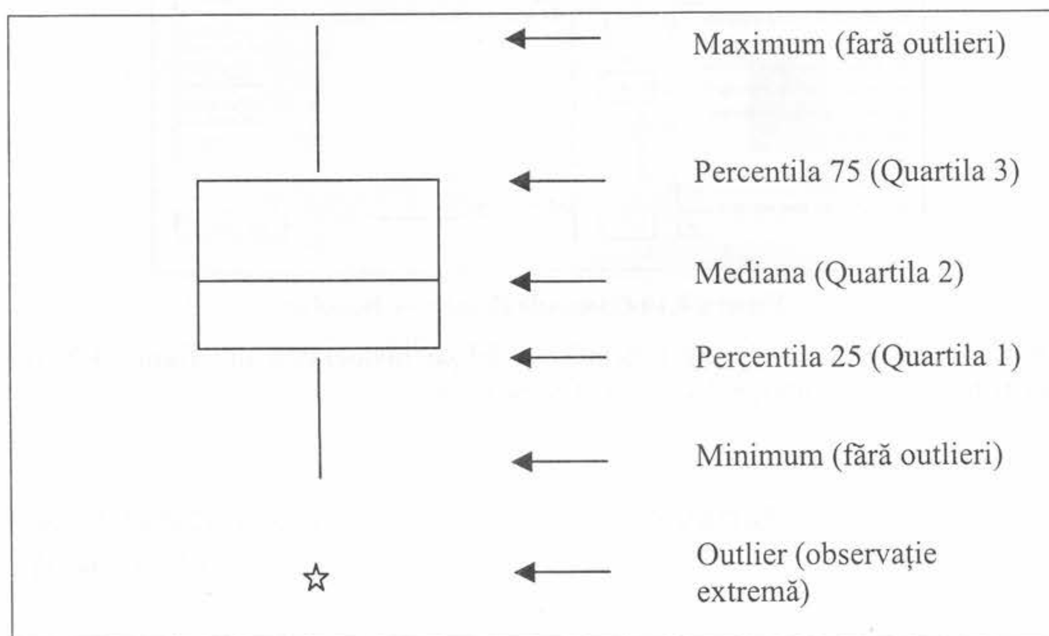


Figura 4.9 *Elementele unei diagrame* Boxplot

Exemplificăm construcţia diagramei *Boxplot* în SPSS pe baza distribuţiei după vârstă, considerată anterior. Demersul de urmat, la fel ca la histogramă, poate fi realizat prin comanda *Boxplot* din meniul *Graphs* (vezi figura 4.10) sau selectând succesiv: meniul *Analyze* → comanda *Descriptive Statistics* →

opțiunea *Explore* → butonul de comandă *Plots* → fereastra *Explore-Plots* (vezi figura 4.11).

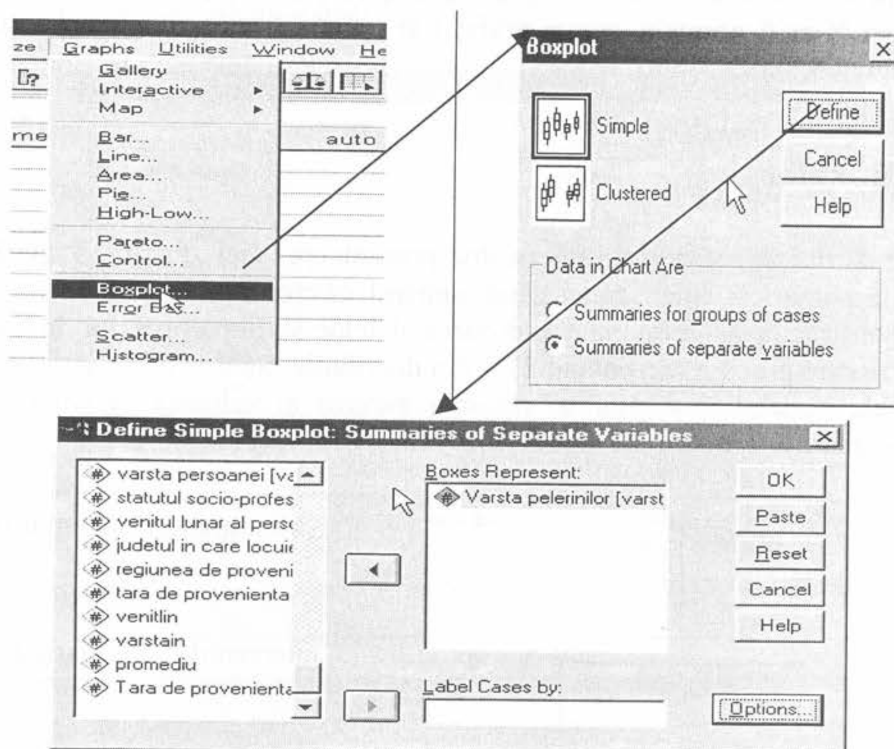


Figura 4.10 *Comanda* Graphs → Boxplot

Boxplot din figura 4.11 ne prezintă, la fel ca histograma din figura 4.5, o distribuție relativ omogenă cu asimetrie pozitivă.

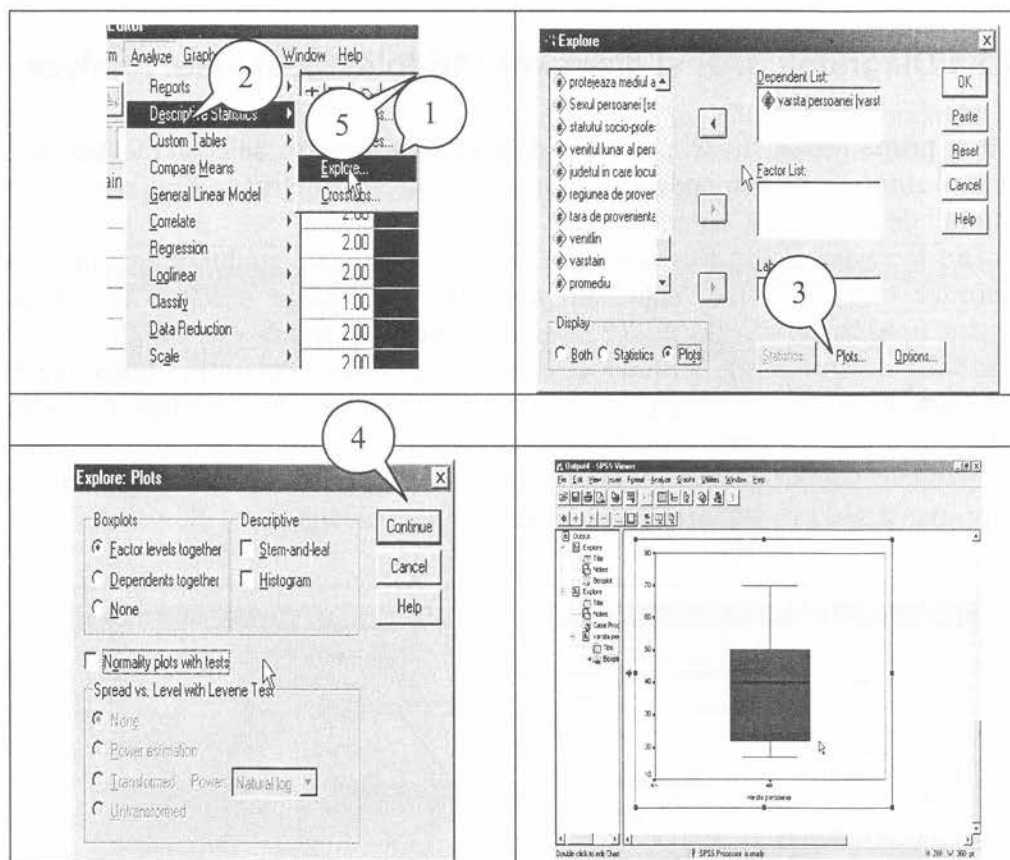


Figura 4.11 *Construirea diagramei* Box plot *prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opţiunea* Explore → Plots → Boxplots

4.3 Grafice pentru distribuţii după o variabilă calitativă (nominală)

Distribuţiile după o variabilă calitativă se prezintă grafic, de regulă, prin diagrame BAR şi PIE.

Diagramele în bare (BAR) şi cercul de structură (PIE) permit să se prezinte frecvenţele la nivelul fiecărei categorii ale unei variabile nominale. Construcţia lor poate fi realizată folosind fie meniul *Analyze*, fie meniul *Graph*.

4.3.1 Diagrama BAR şi diagrama PIE folosind meniul Analyze

Într-o primă variantă, se parcurge demersul în care se selectează succesiv: meniul *Analyze* → comanda *Descriptive Statistics* → opţiunea *Frequencies* → butonul de comandă *Charts*.

Din fereastra dialog *Frequencies Charts*, se stabileşte modul de exprimare a valorilor variabilei (frecvenţe sau procente) şi se alege butonul de opţiuni pentru tipul de grafic dorit: *Bar charts* (pentru bare) sau *Pie charts* (pentru diagramă de structură). Butonul de comandă *Continue* determină revenirea la fereastra *Frequencies*, din care se activează butonul de comandă *OK* care finalizează crearea graficului. Diagrama aleasă se obţine automat în fereastra de rezultate *Output Viewer*, putând fi modificată, prin *Chart Editor*, tipărită la imprimantă sau salvată într-un document Word (vezi figura 4.12).

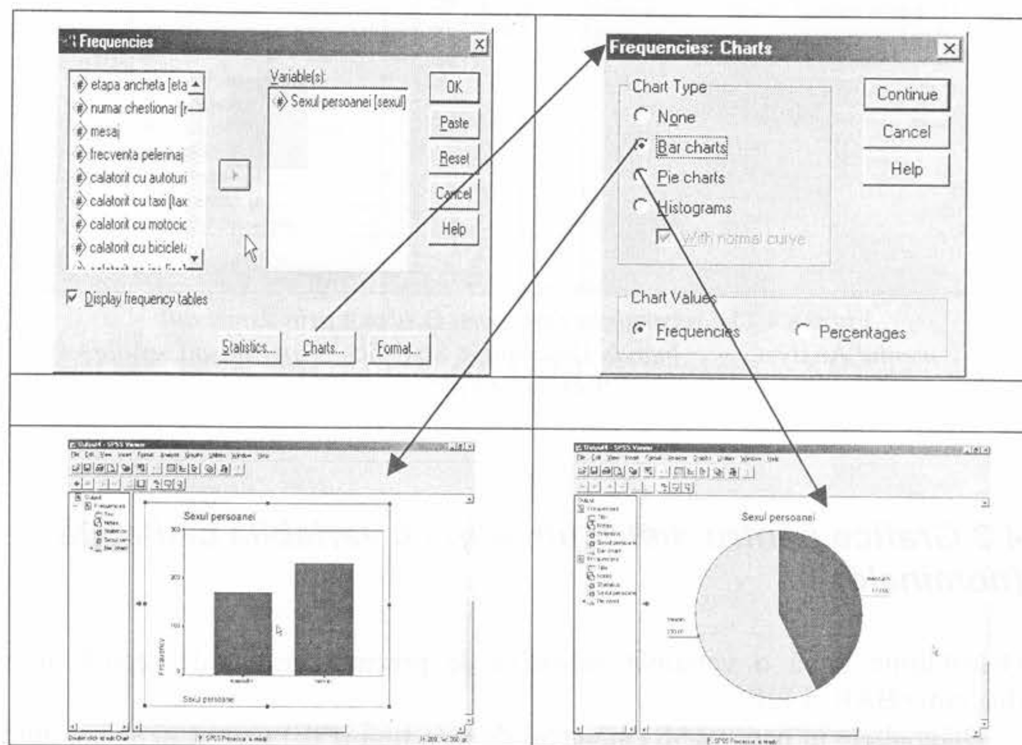


Figura 4.12 *Construirea diagramelor* Bar charts *şi* Pie charts *prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opţiunea* Frequencies → Charts

4.3.2 Diagrama BAR şi diagrama PIE folosind meniul Graph

Cea de-a doua variantă presupune demersul: meniul *Graphs* → comanda *Bar* sau *Pie* → opțiunea *Define Bar* (sau *Define Pie*) → *Simple Bar* (sau *Pie*) *for Groups of Cases* → butonul de comandă *OK* (vezi figura 4.13).

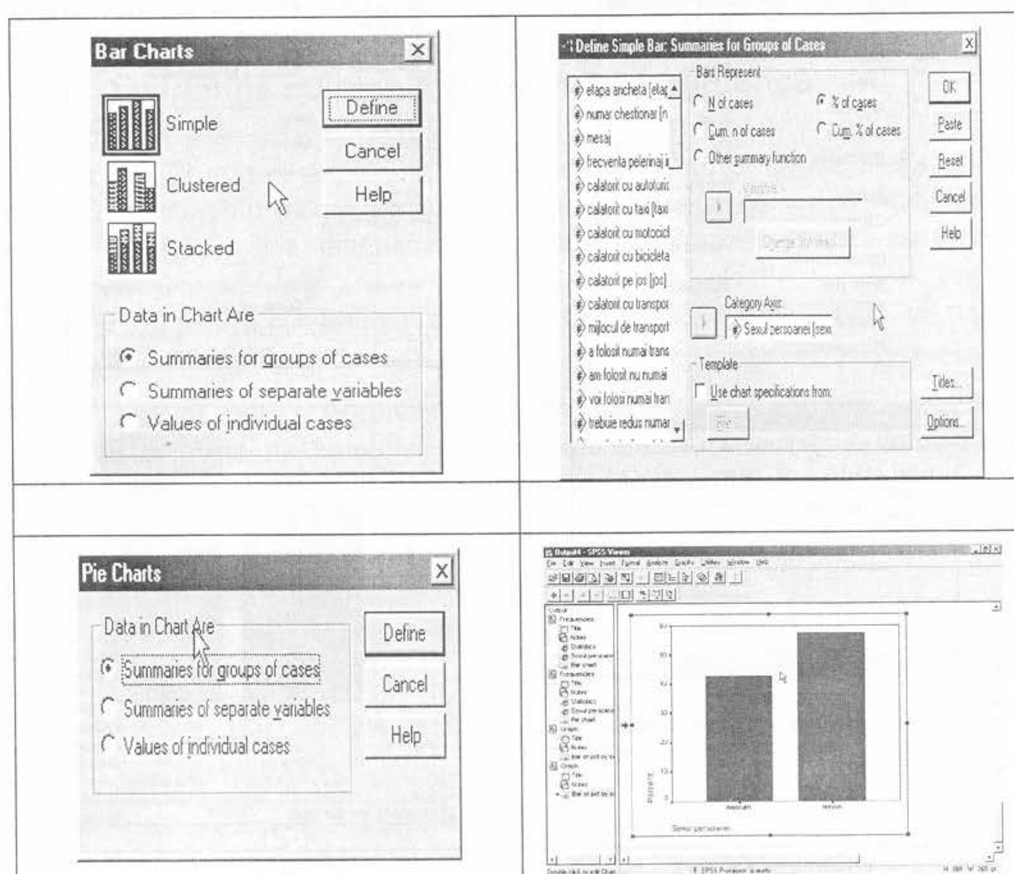


Figura 4.13 *Construirea diagramelor* Bar charts şi Pie charts *prin demersul: meniul* Graphs → *comanda* Bar *sau* Pie

O altă modalitate de construire a acestor diagrame o oferă comanda *Interactive* din meniul *Graphs*. De exemplu, pentru obținerea unei diagrame *Bar*, se selectează opțiunea *Bar* care deschide fereastra *Create Bar Chart*. În această fereastră, în pagina *Assign Variables*, se selectează, prin tragere, variabila categorială, iar în pagina *Bar Chart Options* se stabileşte forma barelor (*Bar Shape*) şi se precizează etichetele acestora (*Count* şi/sau *Value*).

Butonul de comandă *OK* salvează setările şi creează diagrama în fereastra de rezultate *Output Viewer* (vezi figura 4.14).

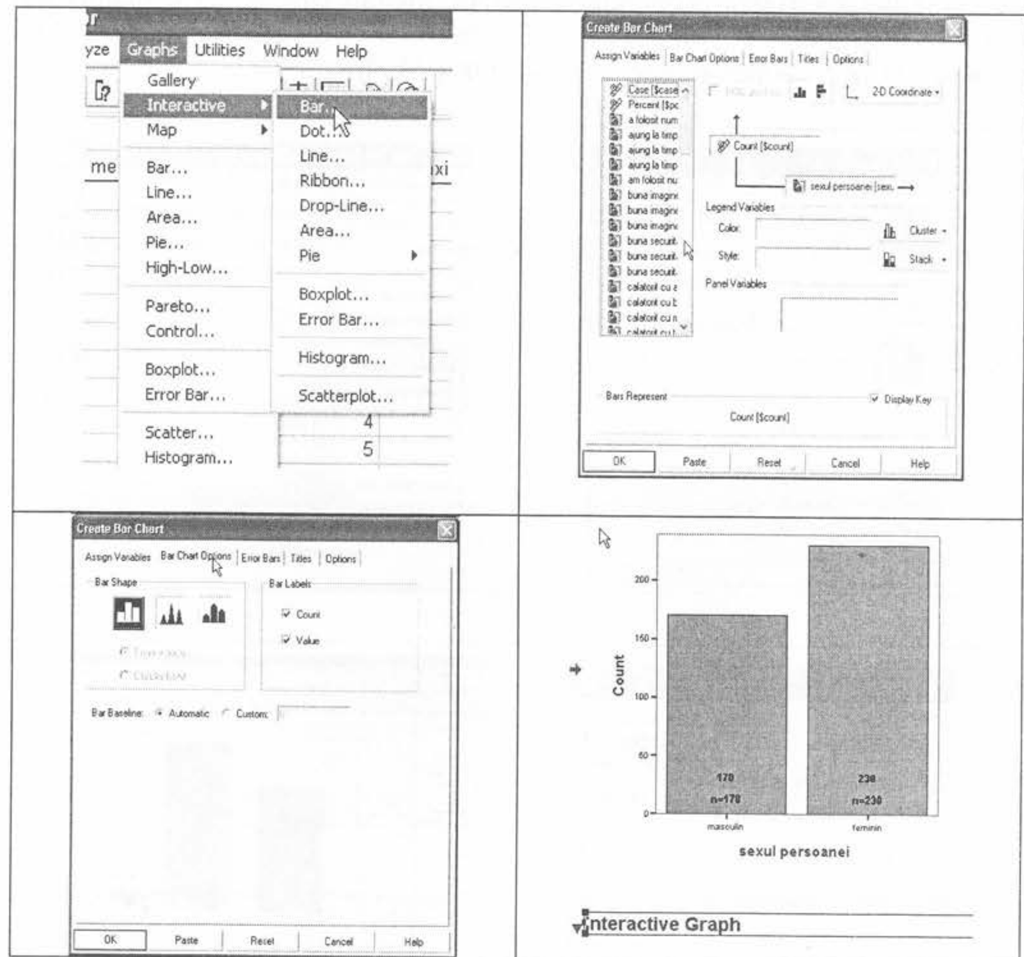


Figura 4.14 *Construirea diagramei* Bar chart *prin demersul: meniul* Graphs → *comanda* Interactive → *opţiunea* Bar

Diagrama de structură (*Pie chart*) şi *Diagrama în bare* (*Bar chart*) reprezintă o cale de a sintetiza un set de date nominale (categoriale). Aşa cum se observă în figura 4.12, *Pie* este un cerc divizat în sectoare. Fiecare sector de cerc reprezintă o categorie, aria acestuia fiind proporţională cu numărul de cazuri din această categorie a variabilei nominale. Diagrama *Bar* este adesea folosită pentru a ilustra categoriile unei distribuţii într-o formă convenabilă. Diagrama prezintă atâtea bare câte categorii are o variabilă. Barele au aceeaşi bază, egală

cu unitatea, iar înălţimea proporţională cu frecvenţa categoriei, astfel încât aria fiecărei bare reprezintă numărul cazurilor categoriei considerate. De exemplu, grupa persoanelor de sex masculin, prezentată în figura 4.14, este formată din 170 de persoane.

4.4 Grafice pentru distribuţii bivariate

4.4.1 O variabilă nominală şi o variabilă numerică

Reprezentarea grafică simultană a unei variabile nominale şi a unei variabile numerice este folosită pentru prezentarea mediilor şi abaterilor standard pe grupe (categorii). Ca diagrame, sunt alese următoarele tipuri: *Histogram*, *Boxplots*, *Stem-and-leaf*.

În SPSS, pentru construirea unor astfel de diagrame, sunt urmate, de regulă, două căi:

a. Meniul *Analyze*→ comanda *Descriptive Statistics*→ opţiunea *Explore*.

Se mută variabila numerică în *Dependent List* şi variabila nominală în *Factor List*. Se alege tipul diagramei dorite (*Histogram*, *Boxplots*, *Stem-and-leaf*.) (vezi figura 4.15).

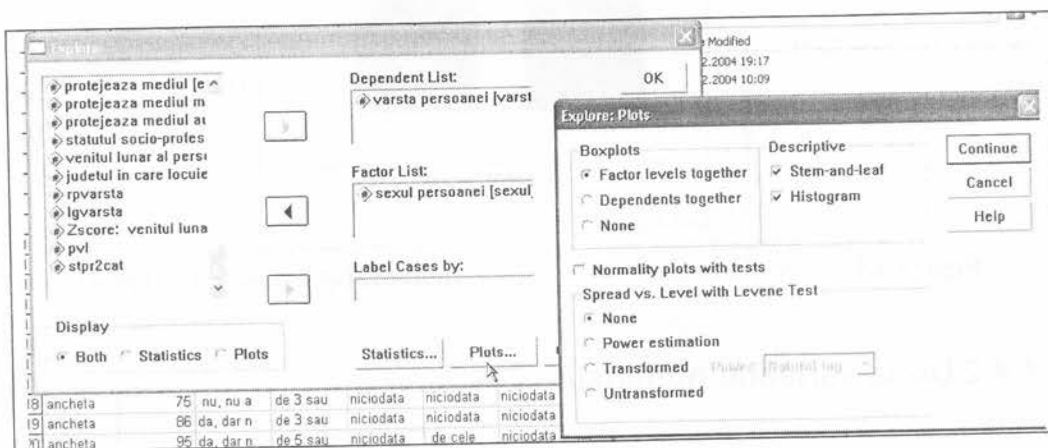


Figura 4.15 *Construirea graficelor prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opţiunea* Explore

b. Meniul *Graphs* → comanda *Interactive* → opţiunea *Histogram*.

Se mută, prin tragere, variabila numerică pe axa abscisei, iar variabila nominală în zona *Panel Variables*.

Diagramele pentru distribuţia după vârstă şi sex sunt prezentate în figurile 4.16 şi 4.17.

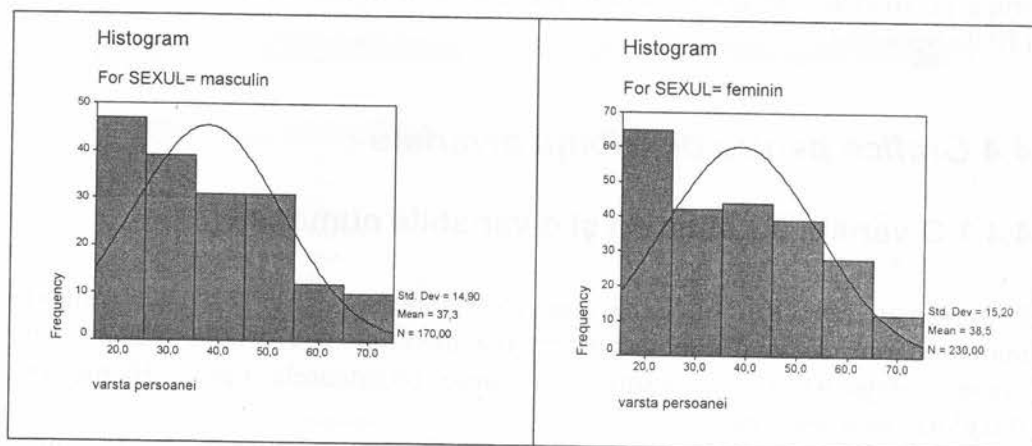


Figura 4.16 *Reprezentarea distribuţiei după vârstă şi sex, folosind histograme*

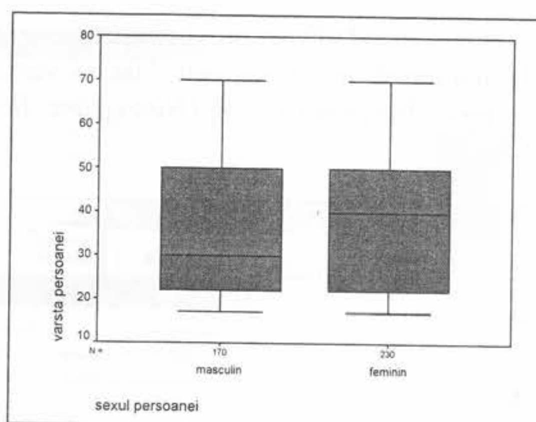


Figura 4.17 *Reprezentarea distribuţiei după vârstă şi sex, folosind box plots*

4.4.2 Două variabile nominale

Reprezentarea grafică a două variabile nominale este folosită pentru prezentarea proporţiilor pe grupe (categorii). În acest scop, sunt alese histogramele cu un panel de variabile. Construcţia lor presupune următorul demers: meniul *Graphs* → comanda *Interactive* → opţiunea *Pie* → *Clustered*.

Output-ul obţinut este prezentat în figura 4.18.

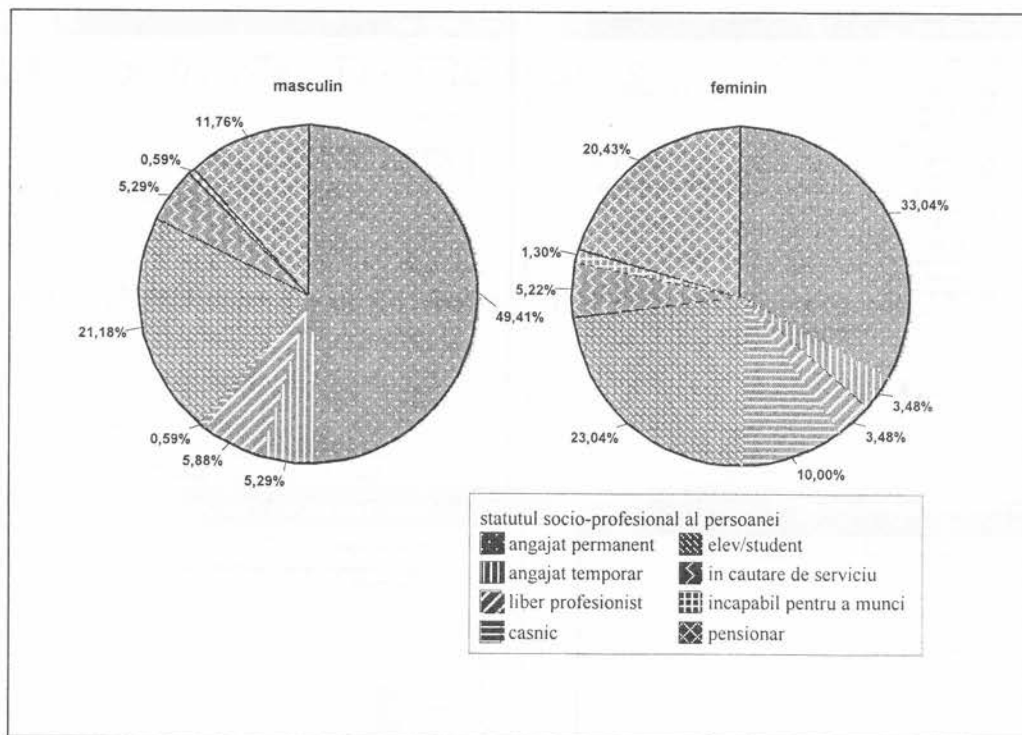


Figura 4.18. *Distribuţia pe sex şi statut socio-profesional, folosind diagrama* Pie

4.4.3 Două variabile numerice

Reprezentarea grafică simultană a două variabile numerice este folosită pentru prezentarea legăturilor dintre fenomene. Ca diagramă, este alesă *Scatterplot.*

Demersul urmat în SPSS pentru a construi *scatterplot* este: meniul *Graphs* → comanda *Interactive* → opţiunea *Scatterplot.*

În fereastra dialog *Create Scatterplot*, în pagina *Fit*, bifăm *Regression*, iar în pagina *Spikes*, bifăm *Fit Line*. Prin comanda *OK*, se obţine în SPSS *Viewer scatterplot*, cu linia de regresie.

Exemplificăm, folosind datele *dez_reg.sav*, pentru variabilele numerice câştigul salarial nominal şi investiţiile pe regiuni, România, anul 2002 (vezi figura 4.19).

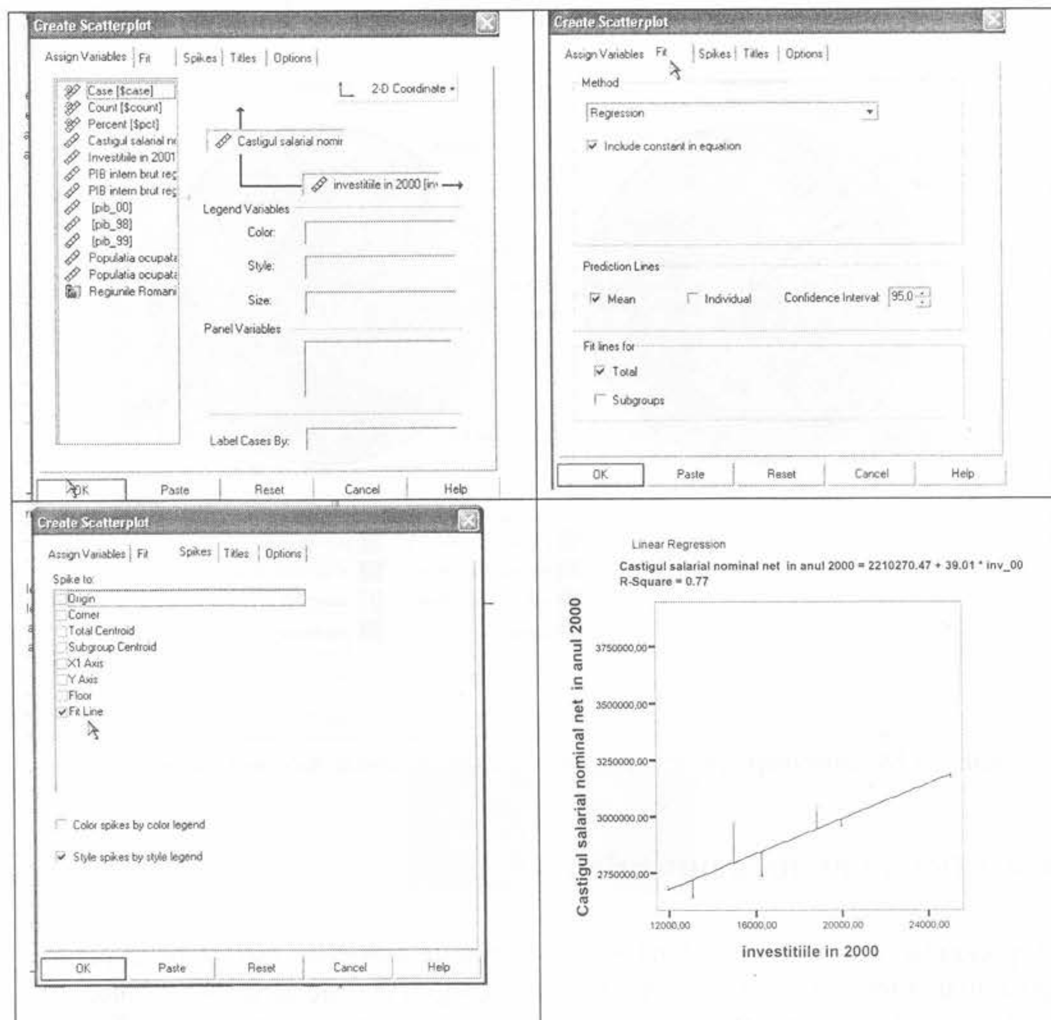


Figura 4.19 *Demersul pentru* Scatterplot, *cu linia de regresie*

Diagrama din figura 4.19 arată că între cele două variabile considerate (investițiile și câștigul salarial nominal) există o legătură liniară, directă, relativ strânsă.

4.5 Modificarea unui grafic în SPSS

Modificarea unui grafic în SPSS poate viza orice element al graficului şi se efectuează prin *Chart Editor*.

4.5.1 Modificarea numărului de intervale pe axa abscisei

Pentru a schimba numărul de intervale pe axa abscisei, se efectuează demersul prezentat în figura 4.20.

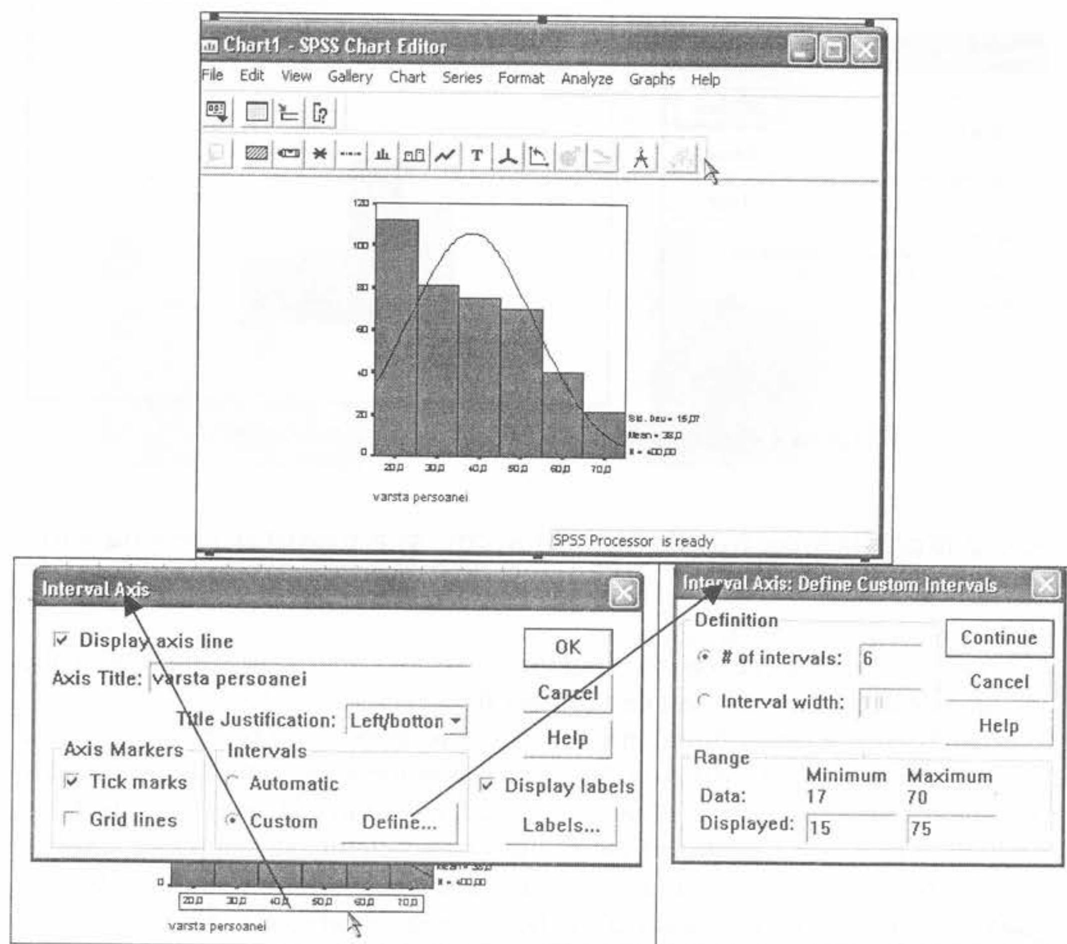


Figura 4.20 *Modificarea numărului de intervale pe abscisa histogramei*

Prin dublu clic pe histogramă, se deschide *Chart Editor*. Apoi, dublu clic pe numerele de sub axa abscisei deschide fereastra *Interval Axis*. În această fereastră, se selectează *Custom* şi se activează, prin clic simplu, butonul de comandă *Define*. Se deschide fereastra de dialog *Interval Axis: Define Custom Intervals*, în care selectăm opţiunea *# of intervals* şi scriem numărul de intervale dorit. În exemplul dat, erau 6 intervale şi le schimbăm cu 4.

Clic pe butonul de comandă *Continue* determină revenirea în fereastra *Interval Axis* din care, prin butonul *OK*, comandăm în SPSS obţinerea histogramei cu un număr de intervale schimbat (vezi figura 4.21). După efectuarea modificării dorite, se închide fereastra *Chart Editor*.

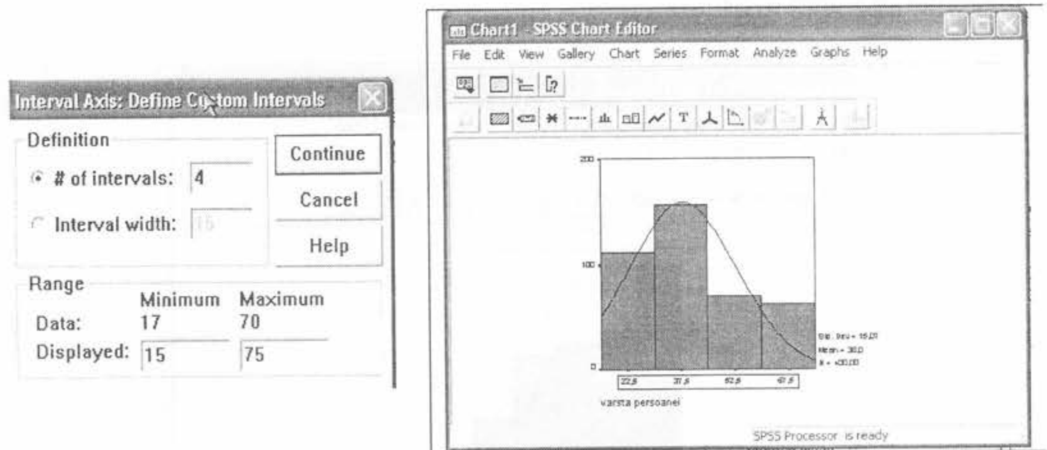


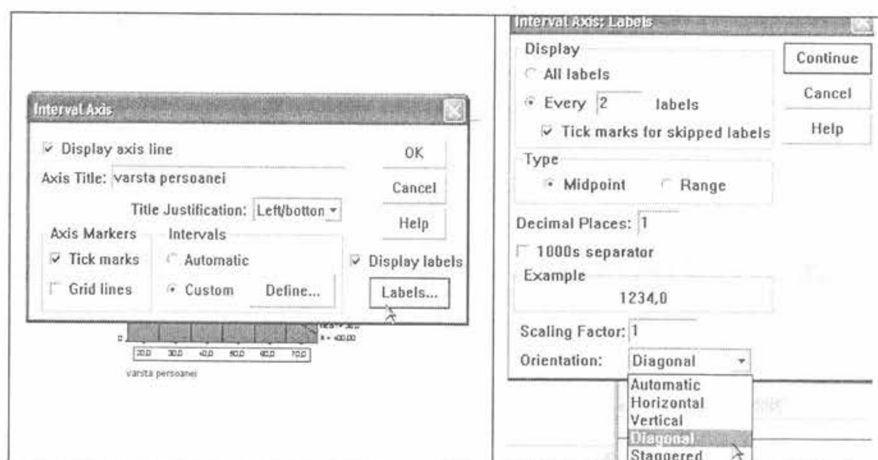
Figura 4.21 *Histograma modificată cu număr de intervale*

4.5.2 Modificarea numărului de spaţii şi a orientării etichetelor de pe axa abscisei

Această operaţie începe, ca orice modificare asupra unui grafic, prin dublu clic pe diagramă, care are ca efect deschiderea ferestrei *Chart Editor*.

În fereastra Chart Editor, prin clic pe etichetele de sub axa abscisei, se deschide fereastra *Interval Axis*. Se activează butonul de comandă *Labels* care deschide fereastra *Interval Axis Labels*, unde în zona *Display* se selectează opţiunea *All labels* (pentru afişarea tuturor etichetelor) sau opţiunea *Every... labels*, pentru a preciza raţia de afişare a etichetelor. În exemplul dat, s-a stabilit pasul 2, ceea ce înseamnă că etichetele se vor afişa din două în două. În aceeaşi fereastră, se selectează, din lista *Orientation*, opţiunea dorită, pentru

modul în care vor fi orientate etichetele: pe orizontală, verticală, diagonală etc. (vezi figura 4.22).



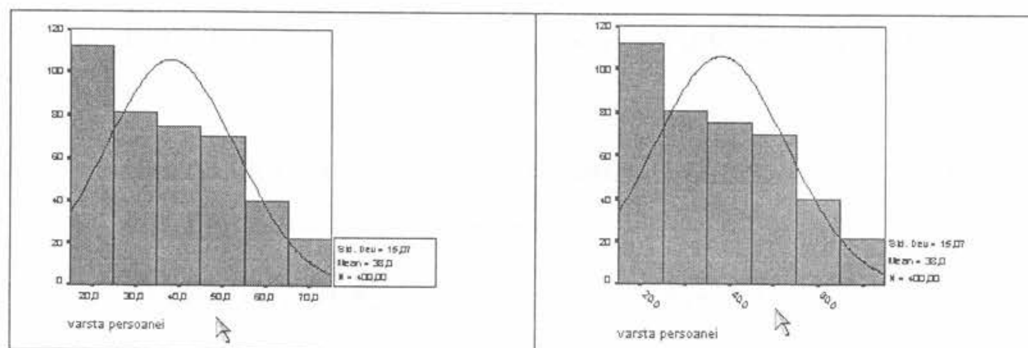


Figura 4.22 *Schimbarea raţiei de afişare a etichetelor şi a orientării acestora*

Alte modificări asupra elementelor unui grafic se pot realiza aplicând un demers asemănător.

CAPITOLUL 5

PARAMETRII UNEI DISTRIBUȚII STATISTICE

- **Indicatori ai tendinței centrale, dispersiei și formei unei distribuții statistice univariate**
- **Calculul indicatorilor tendinței centrale, dispersiei și formei unei distribuții univariate în SPSS**
- **Parametrii unei distribuții bivariate (bidimensionale)**
- **Calculul parametrilor unei distribuții bivariate folosind SPSS**

În acest capitol, sunt tratați indicatorii folosiți pentru caracterizarea unei distribuții statistice. Va fi prezentat modul de calcul (manual și în SPSS), precum și modul de interpretare a indicatorilor sintetici descriptivi la nivelul unui eșantion și al unei populații.

Prezentarea indicatorilor se face distinct pentru distribuții univariate și distribuții bivariate, ținând cont de natura variabilelor și modul lor de măsurare.

5.1 Indicatori ai tendinței centrale, dispersiei și formei unei distribuții statistice univariate

Datele statistice prezentate într-un tabel de frecvență pot fi rezumate cu ajutorul indicatorilor tendinței centrale, dispersiei și formei unei distribuții.

5.1.1 Indicatori ai tendinței centrale

Indicatorii tendinței centrale exprimă în mod sintetic și generalizant ceea ce este normal într-o distribuție din punctul de vedere al unei variabile statistice. În jurul lor se grupează celelalte valori observate. Indicatorii tendinței centrale sunt prezentați pe tipuri de variabile.

Cazul unei variabile numerice

Media este punctul de echilibru al tuturor valorilor unei distribuții. Este o mărime ușor de calculat. Pentru o variabilă *X*, media se calculează după relațiile:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$, pentru o populație de volum N;

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$, pentru un eșantion de volum n.

Mediana (*Me*) este punctul central al unei distribuții, valoarea care separă ansamblul datelor unei serii ordonate în două părți egale, 50% din observații se găsesc sub această valoare și 50% se află peste această valoare. Aflarea medianei presupune depistarea directă a valorii centrale.

În cazul variabilelor numerice discrete cu frecvenţele egale între ele, calculul este direct, folosind relaţia:

$$Me = x_{\frac{n+1}{2}},\ Me = \frac{x_{n/2} + x_{(n/2)+1}}{2}.$$

În cazul variabilelor continue, aflarea medianei se realizează prin interpolare, folosind relaţia:

$$Me = x_{i-1} + d\frac{U^{Me} - N_{i-1}}{n_i},$$

unde:

$$U^{Me} = \frac{n+1}{2},\ N_{i-1} = N_i - n_i,\ \text{iar}\ N_i = \sum_{h=1}^{i} n_h.$$

Modul (Mo), numit şi *dominantă* (Do), este valoarea cea mai frecventă într-o distribuţie, adică valoarea unei variabile purtată de frecvenţa maximă. Se poate afla direct, prin citirea valorii x_i corespunzătoare frecvenţei maxime, în cazul unei *variabile numerice discrete*, şi prin interpolare liniară, în cazul unei *variabile continue*, după relaţia:

$$Mo = x_{i-1} + d\frac{\Delta_1}{\Delta_1 + \Delta_2},$$

unde:

$$d = x_i - x_{i-1}, \Delta_1 = n_i - n_{i-1}, \Delta_2 = n_i - n_{i+1}.$$

Cazul unei variabile nominale

Pentru o variabilă nominală (categorială), pot fi calculaţi următorii parametri: proporţia şi modul.

Proporţia este simbolizată prin p_i (pe ansamblul unei populaţii de volum N), respectiv f_i (la nivelul unui eşantion observat, de volum n), cu $i = \overline{1,k}$, unde k reprezintă numărul de categorii. Se calculează ca raportul între parte şi întreg. Pentru a facilita interpretarea, se foloseşte expresia procentuală, înmulţind raportul cu 100. Suma lor este egală cu 1, respectiv 100%.

Modul unei variabile nominale reprezintă categoria cea mai des întâlnită.

Pentru o variabilă nominală ordinală, în plus, se poate calcula şi *mediana*.

Comparaţii între medie, mediană şi mod

Toţi parametrii tendinţei centrale se exprimă în aceleaşi unităţi de măsură ca şi variabila observată.

Media este reprezentativă pentru distribuţii omogene, dar este influenţată de valorile extreme ale variabilei observate şi este nereprezentativă pentru distribuţiile eterogene. Se pretează la calcule algebrice. Media se calculează numai pentru variabile numerice. Nu are sens calculul acesteia pentru variabile nominale. De exemplu, ar fi absurd să se calculeze media celor două categorii, masculin şi feminin, ale variabilei „sexul persoanei".

Mediana are avantajul că nu este influenţată de valorile extreme ale unei serii, dar are dezavantajul că, neţinând seama de ansamblul datelor, este o valoare aproximativă. Mediana se poate calcula şi pentru variabile ordinale.

Modul are avantajul că nu este influenţat de valorile extreme, este ideal pentru populaţii eterogene, dar este o mărime aproximativă, depinzând de alegerea intervalului dominant. Modul se poate calcula şi pentru variabile nominale (categoriale).

Cele trei mărimi medii fundamentale sunt egale între ele în cazul seriilor simetrice, dar sunt inegale în cazul seriilor asimetrice. Media, pe baza celor mai mici pătrate, minimizează suma pătratelor abaterilor între valorile observate şi parametrul tendinţei centrale. Această sumă este întotdeauna inferioară sau egală cu suma pătratelor abaterilor între valorile observate şi mediană sau mod:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 \le \sum_{i=1}^{n}(x_i-\text{Me}) \text{ şi } \sum_{i=1}^{n}(x_i-\bar{x})^2 \le \sum_{i=1}^{n}(x_i-Mo)^2.$$

5.1.2 Quantile

Quantilele sunt mărimi care împart o distribuţie într-un număr de părţi egale. Au acelaşi mod de calcul ca al medianei, aceasta fiind quantila care împarte distribuţia în două părţi egale. Quantilele cele mai des utilizate sunt: quartilele, decilele şi centilele.

Quartilele (Q) sunt în număr de trei şi împart datele în patru părţi egale.

Decilele (D) sunt în număr de nouă şi împart datele în zece părţi egale.

Centilele (C) sunt în număr de 99 şi împart datele în 100 de părţi egale.

Între quantile există relația:

$Me = Q_2 = D_5 = C_{50}$.

Quantilele sunt folosite pentru interpretarea dispersiei.

5.1.3 Indicatori ai dispersiei

Dispersia reprezintă fenomenul de împrăştiere a valorilor individuale x_i ale unei variabile X, față de nivelul lor mediu.

Cazul unei variabile numerice

În cazul variabilelor numerice (tip scală), parametrii dispersiei sunt: amplitudinea variației, varianța, abaterea medie pătratică, abaterea medie liniară, coeficientul de variație.

Amplitudinea variației exprimă diferența dintre valoarea cea mai mare şi valoarea cea mai mică ale unei variabile observate şi se stabileşte după relația:

$A_x = x_{max} - x_{min}$.

Varianța este media pătratelor abaterilor valorilor individuale de la media lor. Este un indicator abstract, folosit pentru calculul abaterii medii pătratice. Varianța se calculează după relațiile:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$, pentru o populație de volum N;

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$, pentru un eşantion de volum n.

Estimația varianței unei populații, calculată pe baza unui eşantion, foloseşte relația:

$$s'^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Observație! Estimarea varianței unei populații pe baza datelor unui eşantion presupune divizarea prin $n - 1$.

Abaterea medie pătratică (deviația standard) măsoară dispersia în jurul mediei şi se calcuează ca radăcină pătrată din varianță:

$$\sigma = \sqrt{\sigma^2}, \text{ respectiv } s = \sqrt{s^2}.$$

Abaterea medie pătratică se măsoară în aceleaşi unități de măsură ca şi media. Cu cât valoarea sa este mai mare în raport cu media, cu atât populația este mai eterogenă, respectiv cu cât valoarea sa este mai mică în raport cu media, cu atât arată o concentrare mai mare a datelor în jurul mediei, populația fiind mai omogenă.

Abaterea medie liniară exprimă distanța medie (în valoare absolută) care separă observațiile individuale față de media lor şi se calculează după relația:

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

Coeficientul de variație este un parametru al dispersiei, calculat în expresie relativă. Se află fie ca raport între abaterea medie pătratică şi medie, fie ca raport între abaterea medie liniară şi medie, după relațiile:

$$v = \frac{s}{\bar{x}} \cdot 100; \quad v = \frac{\bar{d}}{\bar{x}} \cdot 100.$$

Pentru facilitarea interpretării, raportul se multiplică cu 100, exprimându-se procentual. Coeficientul de variație este, astfel, independent de unitatea de măsură.

Cazul unei variabile nominale

În cazul variabilelor nominale, pentru măsurarea dispersiei, se calculează *indicatorii diversificării*, cel mai cunoscut fiind *indicele de diversificare.*

Indicele de diversificare este cunoscut în literatura de specialitate[1] şi sub denumirea de *valoarea Agresti*, V_s.

Valoarea V_s reprezintă suma probabilităților (p_i) ca două unități statistice dintr-o colectivitate N să facă parte din categorii diferite, k, definite după o variabilă observată, X, la nivelul colectivității:

$$V_s = \sum_{i=1}^{k} p_i \cdot (1 - p_i).$$

1. A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York, 1990, p. 24.

Considerând frecvenţa relativă f_i drept estimator al lui p_i, se poate nota:

$$V_s = \sum_{i}^{k} f_i(1 - f_i) = 1 - \sum_{i=1}^{k} f_i^2 .$$

Valoarea V_s, folosită ca măsură a dispersiei unei serii calitative, poate lua valori în intervalul: $\left[0; 1 - \frac{1}{k}\right]$. Valoarea minimă *(0)* corespunde cazului în care toate unităţile unei colectivităţi aparţin unei singure categorii; valoarea maximă este atinsă în cazul unei distribuţii uniforme (frecvenţe egale) în cele *k* activităţi.

De exemplu, considerând distribuţia populaţiei ocupate pe 11 grupe de activităţi ale economiei naţionale, în România, martie 1995, se obţine: $V_S = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,248 = 0,752$ (*Sursa*: Calculat pe baza datelor din „Ancheta asupra forţei de muncă în Gospodării" [AMIGO], p. 13, martie, 1995, C.N.S., România). Valoarea maximă de $\left(1 - \frac{1}{k}\right) = 0,909$ arată, pentru exemplul considerat, o inegalitate accentuată a grupelor de activităţi după numărul populaţiei ocupate.

5.1.4 Indicatori ai formei unei distribuţii

Pentru aprecierea formei unei distribuţii, se folosesc:
- coeficientul de asimetrie;
- coeficientul de boltire sau aplatizare.

Coeficientul de asimetrie exprimă gradul de dezechilibru al unei distribuţii şi se calculează ca raport între momentul centrat de ordin trei (μ_3) la puterea a doua şi momentul centrat de ordin doi (μ_2) la puterea a treia, după relaţia:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} .$$

Când valoarea coeficientului de asimetrie variază între –1 şi 0 indică prezenţa unei distribuţii asimetrice negative, cu abatere spre stânga, iar când variază între 0 şi 1 indică o distribuţie cu asimetrie pozitivă, cu abatere spre dreapta; când ia valoarea 0 indică prezenţa unei distribuţii simetrice.

O valoare a asimetriei mai mare decât 1 indică o distribuție care diferă semnificativ față de o distribuție normală, distribuție simetrică.

Observație! Deoarece majoritatea testelor statistice presupun o distribuție normală, este bine să se verfice valoarea acestui coeficient. Dacă distribuția nu este normală, se recomandă transformarea datelor sau aplicarea testelor neparametrice, care nu impun restricția de nomalitate a unei distribuții.

Coeficientul de boltire sau aplatizare (kurtosis) se calculează în funcție de momentul centrat de ordin patru (μ_4) și momentul centrat de ordin doi (μ_2), după relația:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3.$$

Kurtosis-ul este o măsură a răspândirii fiecărei observații în jurul unei valori centrale. Pentru o distrubuție normală, valoarea *kurtosis*-ului statistic este 0 și se numește *distribuție mezocurtică*. Atunci când coeficientul este mai mare ca zero, indică o grupare mai puternică a valorilor în jurul valorii centrale, curba este mai boltită decât o distribuție normală și se numește *distribuție leptocurtică*. Atunci când coeficientul este mai mic decât zero, indică o grupare mai slabă în jurul valorii centrale, curba frecvențelor este mai aplatizată și se numește *distribuție platicurtică*.

5.2 Calculul indicatorilor tendinței centrale, dispersiei și formei unei distribuții univariate în SPSS

Calculul indicatorilor tendinței centrale, dispersiei și formei unei distribuții univariate cu ajutorul SPSS poate fi realizat prin mai multe căi. În continuare, prezentăm câteva opțiuni din comenzile meniului *Analyze*.

5.2.1 Calculul indicatorilor tendinței centrale, dispersiei și formei unei distribuții prin opțiunea Descriptives: Options

O primă opțiune de calcul pe care o prezentăm este *Descriptives* din comanda *Descriptive Statistics*, subordonată meniului *Analyze* (vezi figura 5.1).

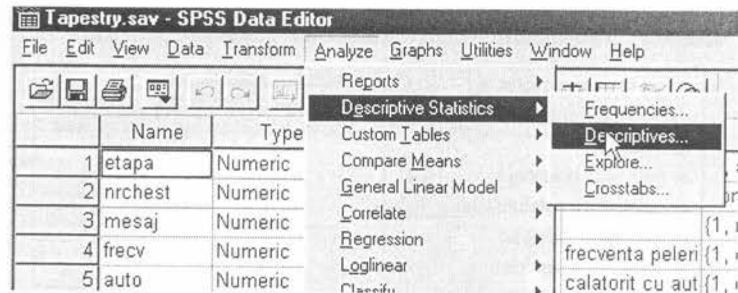


Figura 5.1 *Selectarea opţiunii* Descriptives

După selectarea opţiunii *Descriptives*, se deschide fereastra de dialog *Descriptives* (vezi figura 5.2) care ne permite să selectăm variabila/variabilele pentru care dorim să calculăm parametrii unei distribuţii.

Prin activarea butonului de comandă *Options* din fereastra *Descriptives*, se deschide fereastra de dialog *Descriptives: Options* (vezi figura 5.3). Din această fereastră, selectăm, prin bifare, în caseta/casetele de validare corespunzătoare, indicatorul/indicatorii care urmează a fi calculat(ţi). Se pot realiza următoarele calcule:

- *Mean* (media);
- *Sum* (suma tuturor observaţiilor);
- *Std. Deviation* (abaterea medie pătratică, numită şi abaterea standard);
- *Variance* (varianţa);
- *Range* (amplitudinea variaţiei);
- *Minimum* şi *Maximum* (valoarea minimă şi valoarea maximă a variabilei selectate);
- *S.E. mean* (eroarea medie de selecţie: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$);
- *Kurtosis* (boltirea);
- *Skewness* (asimetria).

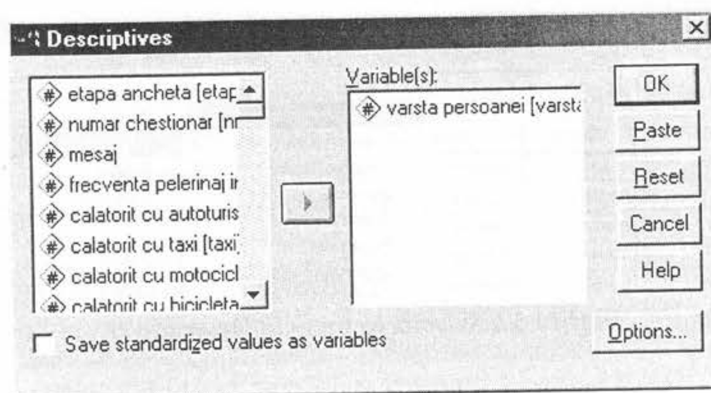


Figura 5.2 *Fereastra* Descriptives

De asemenea, din această fereastră, din zona *Display Order*, se poate alege una din posibilitățile de afișare a rezultatelor (lista variabilelor, ordine crescătoare, ordine descrescătoare etc.).

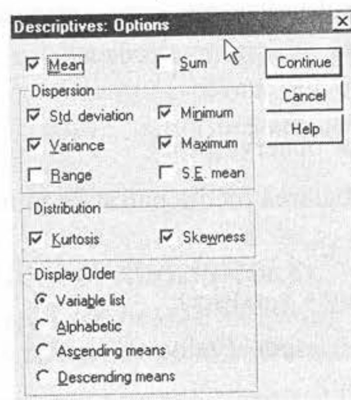


Figura 5.3 *Fereastra de dialog* Descriptives*:* Options

Butonul de comandă *Continue* din fereastra dialog *Descriptives: Options* determină revenirea în fereastra *Descriptives*, din care prin butonul *OK* se comandă obținerea output-ului ce va fi afișat în fereastra de rezultate *Output Viewer*. Pentru exemplificare, folosim baza de date *Tapestry.sav*, rezultatul fiind prezentat în output-ul din figura 5.4.

Output1 - SPSS Viewer

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewnes	Kurtosis
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
varsta persoanei	400	17	70	38.03	15.07	.468	-.822
Valid N (listwise)	400						

Figura 5.4 *Parametrii distribuţiei „ Vârsta pelerinilor" din eşantionul Tapestry-Iaşi, octombrie 2002, calculaţi prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opţiunea* Descriptives

Tabelul de rezultate din output poate fi modificat, de exemplu, prin schimbarea locului statisticilor din coloane cu locul variabilelor din rânduri (vezi paragraful 3.5).

Pentru aceasta, prin dublu clic pe tabelul cu rezultate *Descriptives Statistics* apare fereastra *Output – SPSS Viewer* (vezi figura 5.5), în care selectăm comanda *Pivoting Trays*, din meniul *Pivot* (vezi figura 5.6).

Output2 - SPSS Viewer

File Edit View Insert Pivot Format Analyze Graphs Utilities Window Help

Bookmarks

Transpose Rows and Columns

Pivoting Trays

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.
	Statistic	Statistic	Statistic	Statistic	Statistic
varsta persoanei	400	17	70	38,03	15,07
Valid N (listwise)	400				

Figura 5.5 *Meniul* Pivot

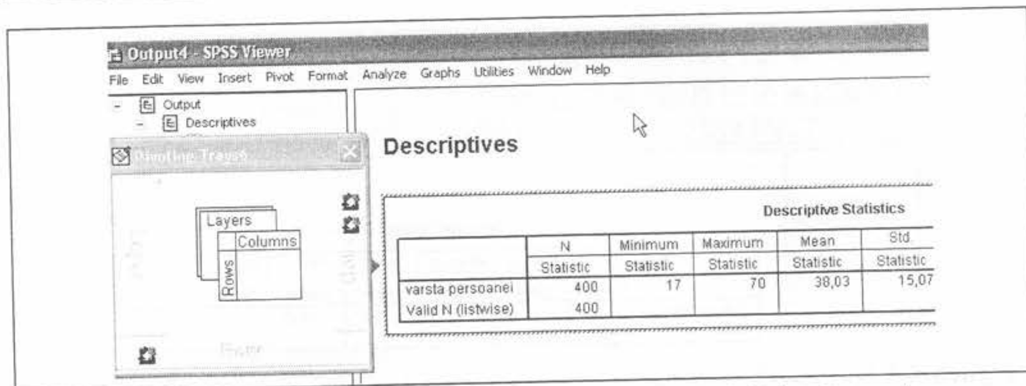


Figura 5.6 *Caseta* Pivoting Trays *de modificare a tabelului de rezultate*

În caseta *Pivoting Trays* schimbăm locul icoanelor floare, trecând cele de pe coloane pe rânduri, respectiv cele de pe rânduri, pe coloane. Prin această operație, se mută variabilele pe coloane și statisticile pe rânduri (vezi figura 5.7).

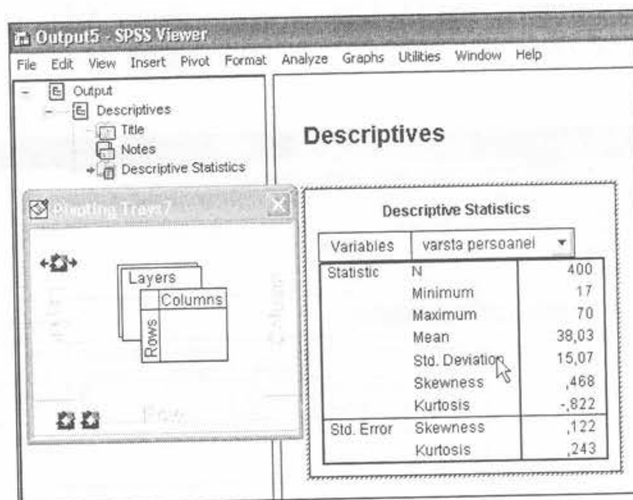


Figura 5.7 *Caseta* Pivoting Trays *cu icoanele schimbate și tabelul modificat*

5.2.2 Calculul indicatorilor statisticii descriptive prin opțiunea Frequencies

O altă cale de obținere a indicatorilor caracteristici ai unei distribuții univariate presupune următoarele selecții succesive: meniul *Analyze*, comanda *Descriptive Statistics*, opțiunea *Frequencies* (vezi figura 5.8).

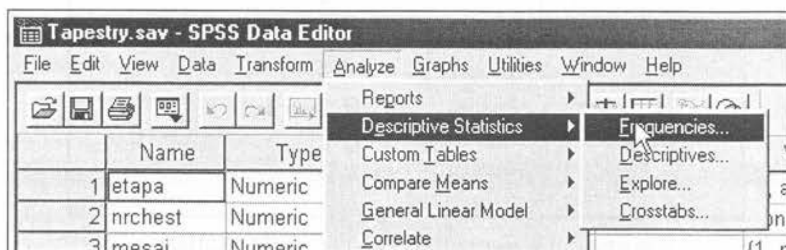


Figura 5.8 *Alegerea opțiunii* Frequencies

Prin selectarea opțiunii *Frequencies*, se deschide fereastra de dialog cu același nume (vezi figura 5.9). În această fereastră, se alege variabila de interes și apoi, prin clic pe butonul de comandă *Statistics*, se deschide fereastra *Frequencies: Statistics* (vezi figura 5.10), din care se pot selecta parametrii doriți, prin bifare în casetele de validare corespunzătoare. Pe această cale se pot afla, pe lângă indicatorii tendinței centrale, indicatorii dispersiei și ai formei unei distribuții, precum și *quartilele* și *centilele* (percentilele).

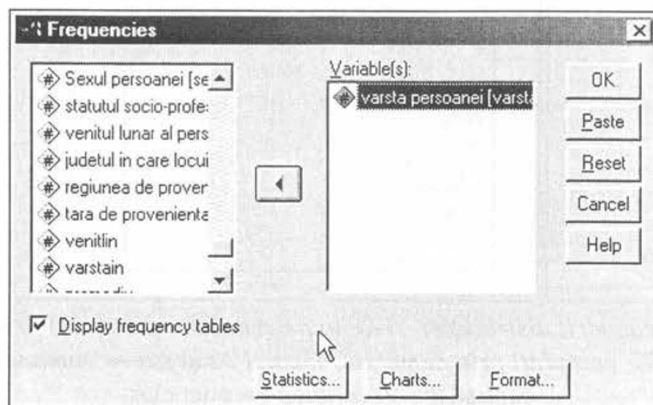


Figura 5.9 *Fereastra* Frequencies

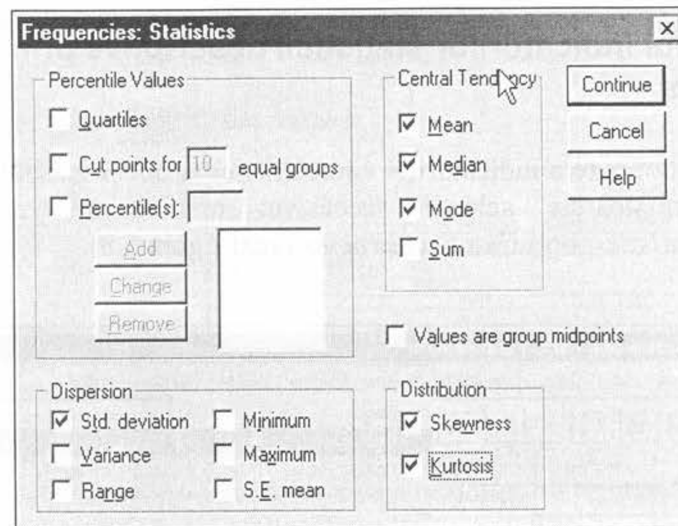
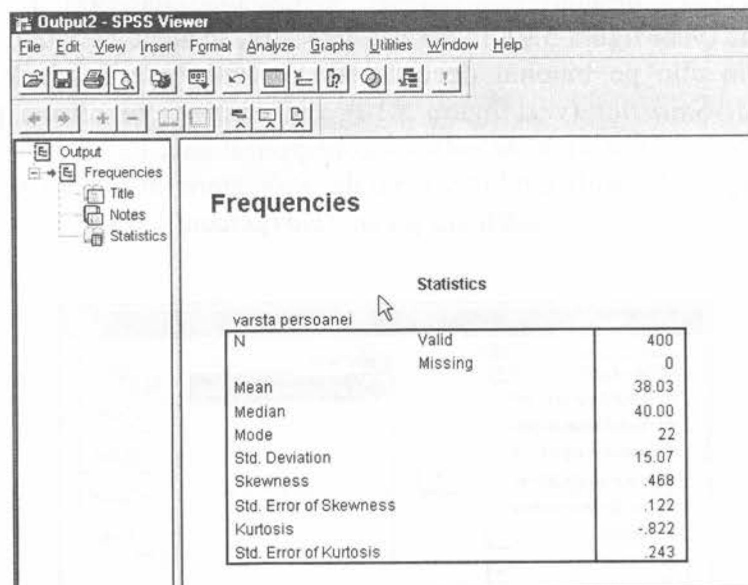


Figura 5.10 *Fereastra* Frequencies: Statistics

Output-ul pentru exemplul considerat anterior este prezentat în figura 5.11.



Statistics

varsta persoanei

N	Valid	400
	Missing	0
Mean		38.03
Median		40.00
Mode		22
Std. Deviation		15.07
Skewness		.468
Std. Error of Skewness		.122
Kurtosis		-.822
Std. Error of Kurtosis		.243

Figura 5.11 *Parametrii distribuției „Vârsta pelerinilor" din eșantionul Tapestry-Iași, octombrie 2002, calculați prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opțiunea* Frequencies

5.2.3 Calculul indicatorilor statisticii descriptive prin opțiunea Case Summaries

O a treia cale de calcul al indicatorilor tendinței centrale, dispersiei și formei unei distribuții univariate, folosind SPSS, este posibilă prin selectarea opțiunii *Case Summaries* din meniul *Analyze*, comanda *Reports*. Această opțiune deschide fereastra *Summary Report: Statistics*, de unde se pot selecta parametrii doriți (vezi figura 5.12).

Output-ul cu indicatorii selectați, pentru același exemplu considerat în paragrafele 5.2.1 și 5.2.2, este prezentat în figura 5.13.

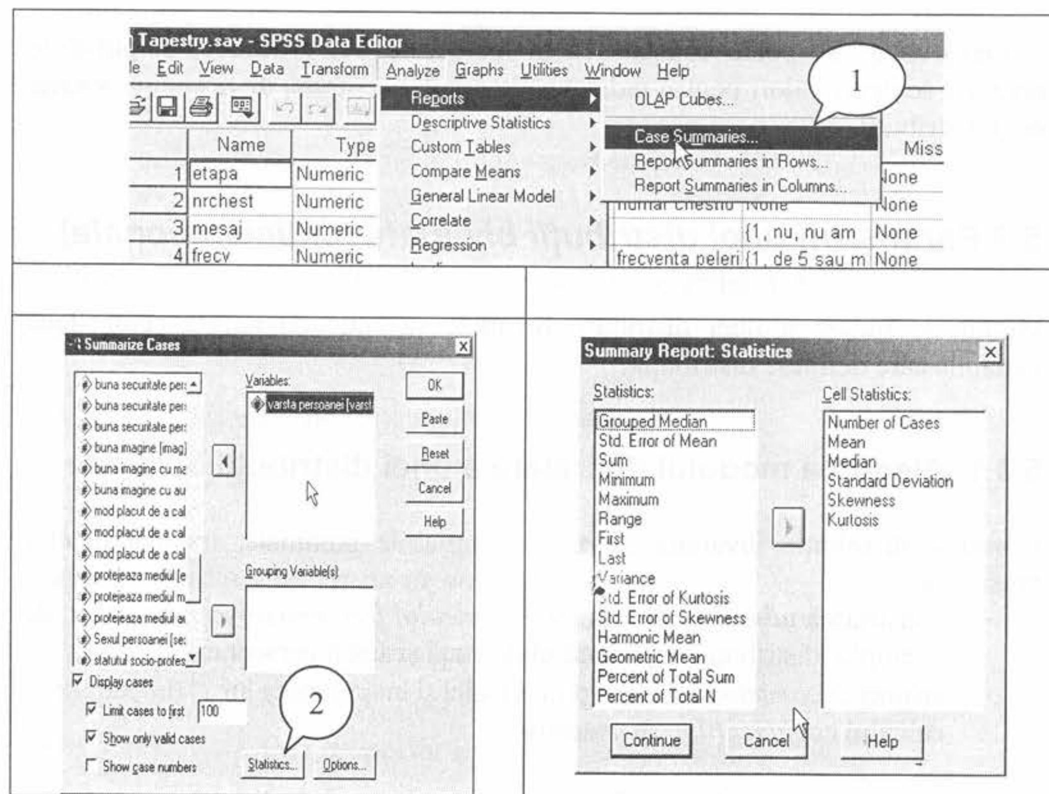
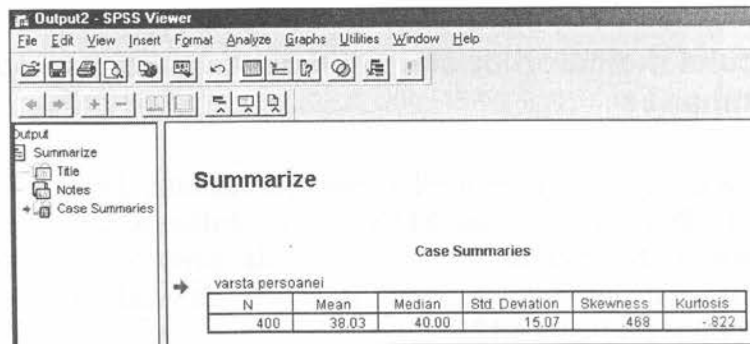


Figura 5.12 *Alegerea indicatorilor unei distribuții univariate prin opțiunea* Case Summaries



N	Mean	Median	Std. Deviation	Skewness	Kurtosis
400	38.03	40.00	15.07	.468	-.822

Figura 5.13 *Parametrii distribuției „Vârsta pelerinilor" din eșantionul Tapestry-Iași, octombrie 2002, calculați prin demersul: meniul* Analyze → *comanda* Reports → *opțiunea* Case Summaries

Observație! Se poate constata că în toate cele trei procedee, output-urile prezintă aceleași valori pentru indicatorii tendinței centrale, dispersiei și formei unei distribuții.

5.3 Parametrii unei distribuții bivariate (bidimensionale)

Modul de tratare a unei distribuții bivariate depinde de natura celor două variabile care definesc distribuția.

5.3.1. Alegerea modului de tratare a unei distribuții bivariate

Pentru o distribuție bivariată cu ambele variabile nominale, tratarea datelor presupune:

- construirea *tabelelor de asociere* și *calculul frecvențelor condiționate* (de exemplu, distribuția după mediul de viață și sexul persoanei);
- analiza *diferențelor calitative* prin calculul și interpretarea lui χ^2 (hi-pătrat);
- calculul *coeficienților de asociere.*

Pentru o distribuție bivariată cu variabile de natură diferită, o variabilă nominală și una exprimată cantitativ, sunt aplicabile:

- procedeul *indicatorilor factoriali ai dispersiei*;
- analiza variației prin *ANOVA*. Procedeul ANOVA măsoară impactul valorilor unor variabile nominale asupra dispersiei valorilor unei variabile cantitative.

Pentru o distribuţie bivariată cu ambele variabile cantitative, sunt aplicabile:

- procedeele folosite în cazul anterior şi, în plus,
- procedeele de determinare a tendinţei centrale şi a dispersiei (medii şi variante condiţionate);
- procedeele de tratare a legăturii dintre variabile (covarianţă, corelaţie, regresie).

5.3.2 Medii şi varianţe condiţionate

Pentru caracterizarea unei distribuţii bivariate cu ambele variabile exprimate cifric se foloseşte un sistem de medii şi varianţe specific: *medii şi varianţe condiţionate, media şi varianţa marginală.*

Medii condiţionate (medii pe grupe)

– Medii condiţionate ale variabilei *X* în raport cu *Y*:

$$\bar{x}_j = \frac{1}{n_{\bullet j}} \cdot \sum_i x_i \cdot n_{ij}, \qquad \text{cu } j = \overline{1, p}.$$

Notaţia $\bar{x}_j$ semnifică media condiţionată a variabilei *X* dacă $Y = y_j$. Se mai notează $\bar{x}_{/y_j}$.

– Medii condiţionate ale variabilei *Y* în raport cu *X*:

$$\bar{y}_i = \frac{1}{n_{i\bullet}} \cdot \sum_j y_j \cdot n_{ij}.$$

Varianţe condiţionate (varianţe de grupă)

– Pentru variabila *X*, condiţionată de $Y = y_j$:

$$\sigma_j^2 = \frac{1}{n_{\bullet j}} \cdot \sum_i (x_i - \bar{x}_j)^2 n_{ij}, \text{ respectiv } \sigma_j^2 = \frac{1}{n_{\bullet j}} \cdot \sum_i x_i^2 n_{ij} - (\bar{x}_j)^2.$$

– Pentru variabila *Y*, condiţionată de $X = x_i$:

$$\sigma_i^2 = \frac{1}{n_{i\bullet}} \cdot \sum_j (y_j - \bar{y}_i)^2 n_{ij}, \text{ respectiv } \sigma_i^2 = \frac{1}{n_{i\bullet}} \cdot \sum_j y_j^2 n_{ij} - (\bar{y}_i)^2.$$

Varianța mediilor condiționate, respectiv *varianța mediilor de grupă* față *de media generală*, δ^2, se calculează după relația:

$$\delta^2 = \frac{1}{\sum_j n_{\bullet j}} \sum_j (\bar{x}_j - \bar{x})^2 \cdot n_{\bullet j},$$

unde:

$\bar{x}_j$ reprezintă *mediile condiționate* ale variabilei X în raport cu Y,

$$\bar{x}_j = \frac{1}{n_{\bullet j}} \cdot \sum_i x_i \cdot n_{ij}, \qquad \text{cu } j = \overline{1,p},$$

$\bar{x}$ *media generală (marginală)*, media pe ansamblul colectivității:

$$\bar{x} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{m} x_i \cdot n_{i\bullet}, \qquad \text{cu } n_{\bullet\bullet} = \sum_{i=1}^{m} n_{i\bullet}.$$

5.3.3 Covarianța

Covarianța a două variabile aleatorii, X și Y, este o măsură a covariației, adică a variației simultane a acestora, și se notează cov(X,Y).

Covarianța este o medie a produselor abaterilor celor două variabile și sintetizează valoarea lor arătând sensul corelației, respectiv al dependenței celor două variabile X, Y.

Calculul covarianței se face după relațiile:

$$cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}, \qquad i = \overline{1,n},$$

respectiv, în cazul datelor prezentate într-un tabel de corelație,

$$cov(X,Y) = \frac{1}{n} \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) n_{ij}, \; i = \overline{1,k}, \; j = \overline{1,p}.$$

Dacă X și Y sunt două variabile aleatorii independente, covarianța este nulă. Reciproca nu este adevărată întotdeauna, adică $cov(X,Y) = 0$ nu implică obligatoriu că X și Y sunt independente.

Apreciere grafică a covarianței. Covarianța poate fi pozitivă sau negativă, conform dispersiei observațiilor în raport cu centrul de greutate, de coordonate $(\bar{x}, \bar{y})$, al norului de puncte.

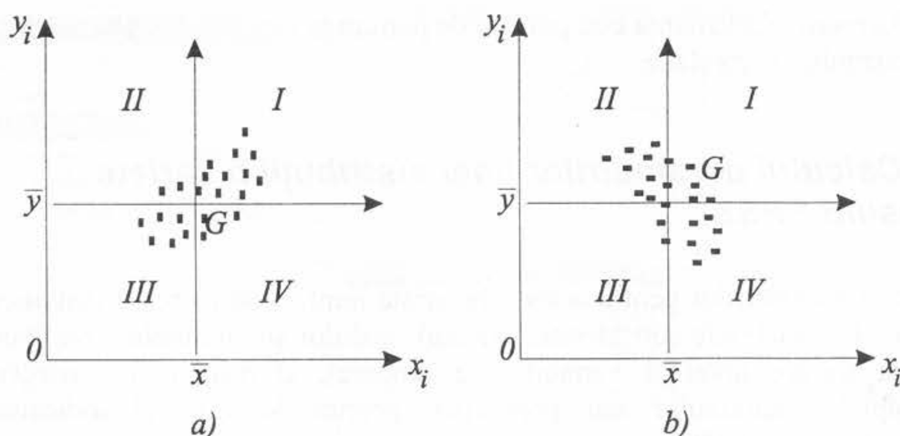


Figura 5.14 *Ilustrarea poziției valorilor* (x_i, y_i) *în raport cu centrul lor de greutate*

În figura 5.14 se observă că, schimbând originea axelor diagramei de dispersie cu centrul de greutate *G* al norului de puncte, valoarea produselor $[(x_i - \bar{x})(y_i - \bar{y})]$ poate fi pozitivă sau negativă. Astfel, se constată că produsele abaterilor punctelor observate sunt, în general, pozitive sau negative (vezi figura 5.14.a sau figura 5.14.b).

Proprietăți ale covarianței. Covarianța are aceleași proprietăți ca varianța cu excepția faptului că valoarea sa poate fi pozitivă sau negativă.

1. *Covarianța este egală cu diferența dintre media produselor și produsul mediilor* celor două variabile:

$$cov(X,Y) = \frac{1}{n}\sum_i x_i y_i - \bar{x} \cdot \bar{y}.$$

Această proprietate facilitează calculul covarianței când mediile au valori zecimale.

2. *Dacă se schimbă originea de calcul al elementelor, covarianța nu se schimbă* dacă și asupra ei se fac aceleași operații:

$cov(X,Y) = d_x d_y\, cov(u,v)$, unde:

$$cov(u,v) = \frac{1}{n}\sum_i (u_i - \bar{u})(v_i - \bar{v}), \text{ în care}$$

$$\bar{u} = \frac{1}{n}\sum_i u_i, \bar{v} = \frac{1}{n}\sum_i v_i, u_i = \frac{x_i - \bar{x}}{\sigma_x}, v_i = \frac{y_i - \bar{y}}{\sigma_y}.$$

Această proprietate ajută la calculul simplificat al covarianței.

Observaţie! Covarianţa este punctul de pornire pentru calculul şi interpretarea coeficientului de corelaţie.

5.4 Calculul parametrilor unei distribuţii bivariate folosind SPSS

Calculul parametrilor pentru o serie bivariată implică obţinerea distribuţiei, în funcţie de variabilele considerate, calculul mediilor şi varianţelor condiţionate (pentru fiecare nivel al variabilei de grupare), al mediei şi varianţei pe ansamblul eşantionului sau populaţiei, precum şi calculul indicatorilor factoriali ai dispersiei.

Pentru exemplificare, folosim baza de date *tapestry.sav*. Considerăm două variabile: *vlunar* – venitul lunar al persoanei (milioane lei) – şi *varsta* – vârsta pelerinilor, recodificată pe grupe (< 25 ani, 25-64 ani, 65 ani şi peste).

5.4.1 Aflarea distribuţiei de frecvenţă bivariate

Distribuţia de frecvenţă „Venitul lunar * Vârsta pelerinilor" exprimă distribuţia eşantionului de persoane observate simultan după cele două variabile considerate, adică arată câte persoane dintr-o anumită categorie de vârstă au un anumit nivel al venitului. Folosind SPSS, distribuţia bivariată se poate obţine pe mai multe căi:

- meniul *Analyze* → comanda *Descriptive Statistics* → opţiunea *Crosstabs*;
- meniul *Analyze* → comanda *Reports* → opţiunea *Case Summaries*;
- meniul *Analyze* → comanda *Descriptive Statistics* → opţiunea *Explore*;
- meniul *Date* → comanda *Split File* → opţiunea *Analyze* → *Reports* → *OLAP Cubes* etc.

Prin demersul *Analyze* → *Descriptive Statistics* → *Crosstabs* se poate obţine o distribuţie bivariată parcurgând paşii prezentaţi în figura 5.15, şi anume:

- se deschide fereastra de dialog *Crosstabs*, în care selectăm variabilele *vlunar* şi *varsta*, din lista variabilelor, şi le mutăm în zonele *Row(s)* şi *Column(s)*;
- din fereastra *Crosstabs*, activând butonul de comandă *Cells*, se deschide fereastra *Crosstabs: Cell Display*, în care bifăm modul dorit de afişare a frecvenţelor în *crosstabel*;
- activarea butonului de comandă *Continue* ne întoarce în fereastra *Crosstabs*, unde prin *OK* se comandă SPSS-ului afişarea output-ului, prezentat în figura 5.16.

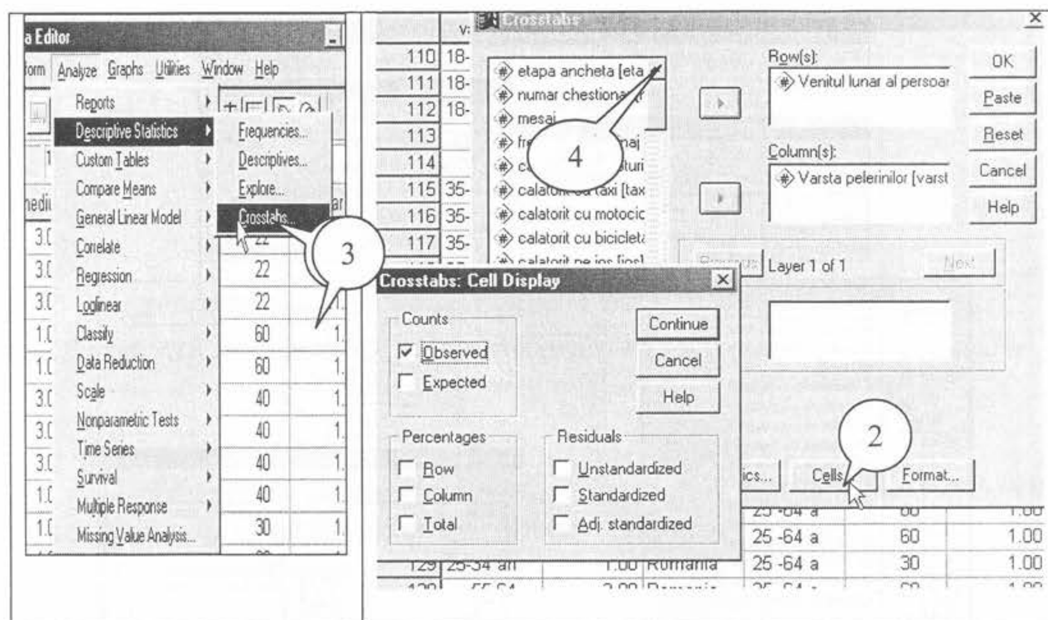


Figura 5.15 *Comenzi pentru obţinerea unui* crosstabel

Venitul lunar al persoanei * Varsta pelerinilor Crosstabulation

Count

		Varsta pelerinilor			
		< 25 ani	25 -64 ani	65 si peste	Total
Venitul lunar al persoanei	1.00	83	102	15	200
	3.00	17	82	7	106
	5.00	5	48		53
	7.00	3	13		16
	9.00	2	9		11
	11.00	2	12		14
Total		112	266	22	400

Figura 5.16 *Distribuţia de frecvenţă „Venitul lunar * Vârsta pelerinilor"*

5.4.2 Calculul mediilor şi varianţelor condiţionate folosind SPSS

Mediile şi varianţele condiţionate se obţin cu ajutorul SPSS parcurgând paşii prezentaţi în figura 5.17, şi anume:

- se selectează meniul *Analyze* → comanda *Reports* → opţiunea *Case Summaries.* Se deschide fereastra de dialog *Summarize Cases*;
- în fereastra *Summarize Cases* selectăm variabilele considerate şi le mutăm prin tragere în zonele *Variables*, respectiv *Grouping Variables*;

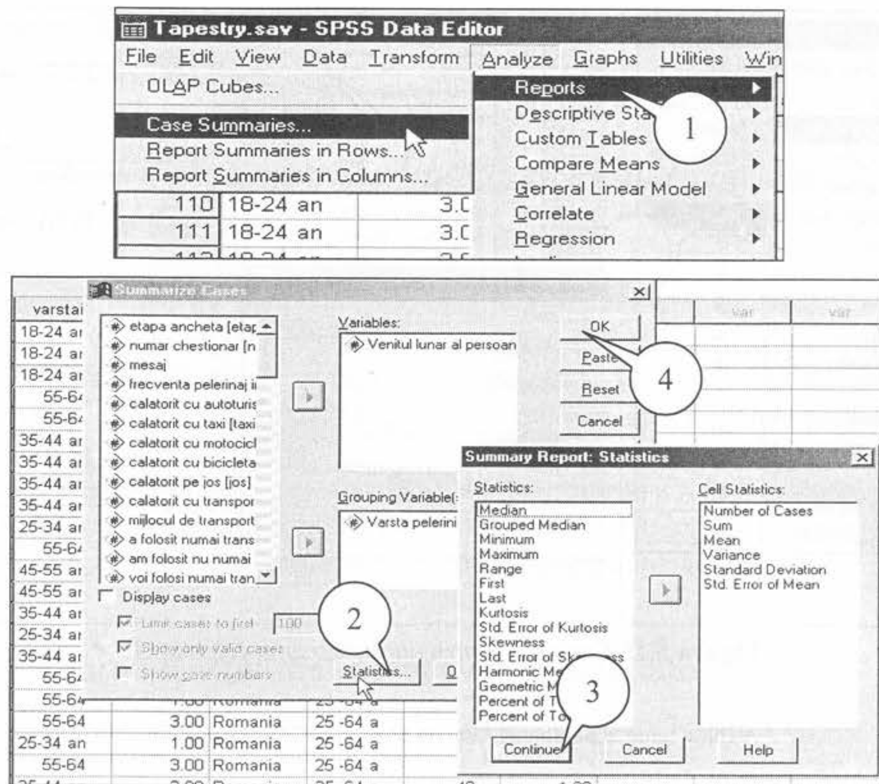


Figura 5.17 *Obținerea mediilor și varianțelor condiționate (pe grupe) și marginale*

- activând butonul de comandă *Statistics* din fereastra *Summarize Cases*, se deschide fereastra *Summary Report: Statistics*, în care selectăm statisticile pe care dorim să le calculăm pentru variabila „Venitul lunar" corespunzător fiecărei categorii de vârstă, adică statisticile condiționate (medii condiționate, varianțe condiționate etc.). Statisticile dorite se selectează din lista *Statistics* și se mută în zona *Cell Statistics*;
- prin clic pe butonul de comandă *Continue*, ne întoarcem în fereastra *Summarize Cases*, unde prin butonul *OK* cerem să se afișeze output-ul cu statisticile dorite pentru variabila „Venitul lunar", corespunzător fiecărei categorii de vârstă (vezi figura 5.18).

Case Summaries

Venitul lunar al persoanei

	Varsta pelerinilor			
	< 25 ani	25 -64 ani	65 si peste	Total
N	112	266	22	400
Sum	220.00	892.00	36.00	1148.00
Mean	1.9643	3.3534	1.6364	2.8700
Variance	4.323	7.044	.909	6.399
Std. Deviation	2.0792	2.6541	.9535	2.5296
Std. Error of Mean	.1965	.1627	.2033	.1265

Figura 5.18 *Medii şi varianţe ale venitului lunar condiţionate de grupa de vârstă a pelerinilor, precum şi valorile marginale ale acestor indicatori*

Interpretarea elementelor din output-ul prezentat în figura 5.18:

N – numărul de persoane intervievate, pe fiecare categorie de vârstă. De exemplu, au fost intervievate 266 de persoane din grupa de vârstă 25-64 ani;

Sum – suma venitului lunar, pe fiecare categorie de vârstă; de exemplu, cele 266 de persoane din grupa de vârstă 25-64 ani au realizat pe ansamblul grupei un venit lunar de 892 milioane de lei;

Mean – media (la nivelul fiecărei categorii de vârstă); de exemplu, oricare din cele 266 de persoane din grupa de vârstă 25-64 ani realizează în medie un venit lunar de 3,3534 milioane lei ;

Variance – varianţa variabilei „Venitul lunar" este calculată pentru fiecare categorie de vârstă;

Std. Deviation – deviaţia standard sau abaterea medie pătratică arată cu cât se abate în medie venitul lunar câştigat de o persoană din grupa de vârstă considerată faţă de venitul lunar mediu al grupei. De exemplu, pentru grupa de vârstă considerată, abaterea medie este de 2,6541 milioane lei, adică aproximativ 68% dintre persoanele din grupa de vârstă 25-64 ani realizează un venit mediu lunar cuprins într-un interval egal cu media grupei plus sau minus valoarea abaterii medii pătratice, respectiv: 3,3534 ± 2,6541 milioane lei;

Std. Error of Mean – eroarea standard a mediei (eroarea medie de reprezentativitate) pentru fiecare grupă de vârstă. Acest indicator este folosit pentru estimarea, prin interval de încredere, a mediei populaţiei, pentru fiecare grupă de vârstă.

Ultima coloană a tabelului din figura 5.18 arată valorile prezentate mai sus corespunzătoare ansamblului eşantionului observat, respectiv *media marginală*, *varianţa marginală* pentru variabila „Venitul lunar" al eşantionului observat.

Aceleaşi rezultate se pot obţine urmând demersul: meniul *Date* → comanda *Split File* → opţiunea *Compare groups*, urmată de meniul *Analyze* → *Reports* →

OLAP Cubes. Acest demers şi output-ul corespunzător sunt prezentate în figura 5.19 şi figura 5.20.

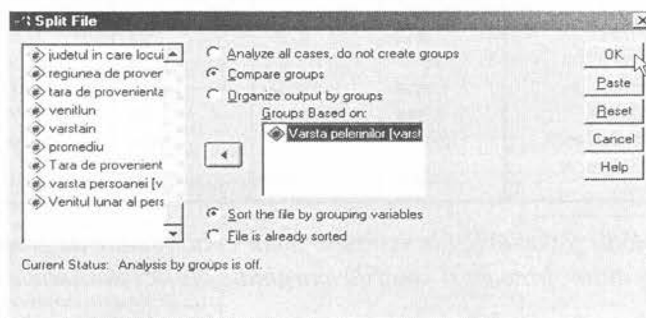


Figura 5.19 *Fereastra* Split File

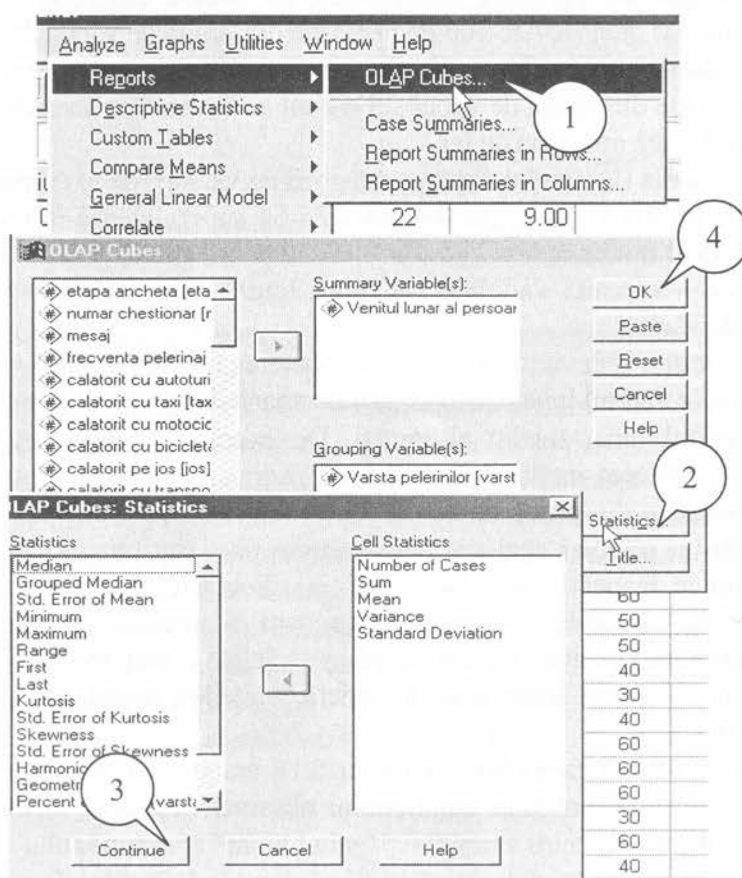


Figura 5.20 *Alegerea statisticilor prin demersul: meniul* Analyze → *comanda* Reports → *opţiunea* OLAP Cubes

5.4.3 Obținerea covarianței folosind SPSS

Covarianța se obține selectând meniul *Analyze* → comanda *Corelate* → opțiunea *Bivariate*. Această opțiune deschide fereastra *Bivariate Corelations* în care activăm butonul de comandă *Options*. Ca efect, se deschide fereastra *Bivariate Correlations: Options*, în care bifăm caseta *Cross-product deviation and covariances* pentru a calcula *covarianța* (vezi figura 5.21).

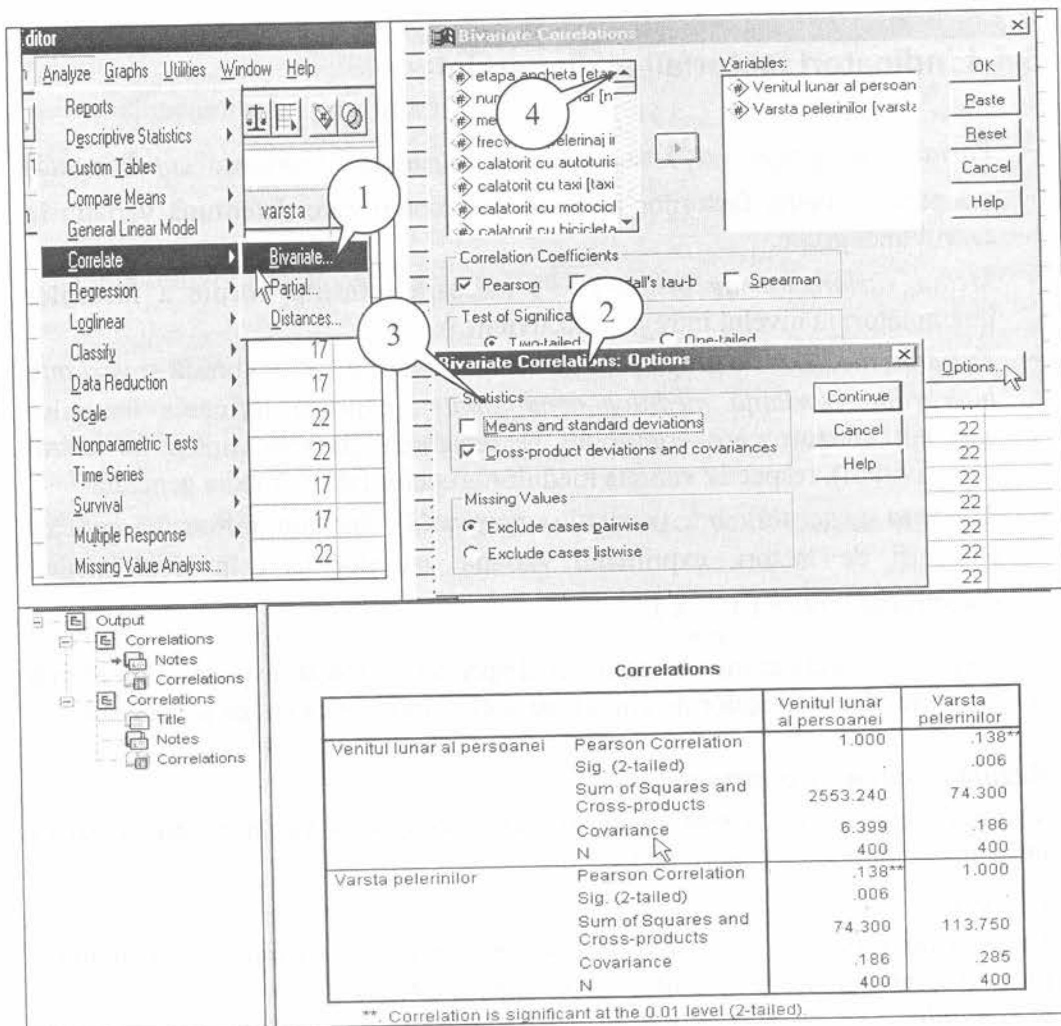


Figura 5.21 *Obținerea covarianței*

Din fereastra *Bivariate Correlations: Options* se activează butonul de comandă *Continue*, care determină revenirea în fereastra *Bivariate Correlations*, din care se selectează butonul *OK* pentru a comanda SPSS calculul covarianţei. Output-ul ne prezintă o *covarianţă* dintre venitul lunar şi vârsta pelerinilor egală cu 0,186 şi o corelaţie *Pearson* egală cu 0,138. Aceste rezultate ne arată că între cele două variabile există o legătură directă, semnificativă la un nivel de risc de 0,01, adică are loc o creştere a veniturilor în funcţie de vârstă, dar legătura este destul de slabă, coeficientul de corelaţie luând o valoare relativ apropiată de zero.

5.4.4 Indicatori factoriali ai dispersiei

- *Varianţa de grupă* (σ_j^2) sau *varianţa intragrupă* (*varianţa condiţionată*) măsoară influenţa factorilor întâmplători, factori care determină variaţia în cadrul unei grupe.
- *Media varianţelor de grupă* ($\overline{\sigma^2}$) măsoară influenţa medie a factorilor întâmplători la nivelul întregii colectivităţi.
- *Varianţa mediilor de grupă faţă de media generală* (δ^2), numită şi *varianţa intergrupe* (*varianţa mediilor condiţionate*), exprimă influenţa factorilor esenţiali (factori care contribuie la separarea grupelor tipice în cadrul colectivităţii), respectiv variaţia mediilor grupelor faţă de media generală.
- *Varianţa generală*, σ^2 (*varianţa marginală*), include influenţa ambelor categorii de factori, exprimând variaţia valorilor (x_i) în jurul mediei colectivităţii totale ($x_i - \bar{x}$).

Observaţie! Indicatorii factoriali ai dispersiei, întrucât exprimă o varianţă (excepţie, media varianţelor de grupă), au acelaşi mod de calcul ca şi varianţa.

Regula de adunare a varianţelor

Varianţa generală (σ^2) este egală cu suma celor două varianţe care măsoară influenţa celor două categorii de factori:

$$\sigma^2 = \overline{\sigma^2} + \delta^2$$

Variaţia sub influenţa factorilor întâmplători şi esenţiali	=	*Variaţia sub influenţa factorilor întâmplători (variaţia reziduală)*	+	*Variaţia sub influenţa factorilor esenţiali*

Pe baza relaţiei de mai sus, se poate afla mărimea oricărei părţi componente a varianţei, după relaţiile:

$\overline{\sigma^2} = \sigma^2 - \delta^2$, respectiv

$\delta^2 = \sigma^2 - \overline{\sigma^2}$.

Coeficienți de măsurare a influenței celor două categorii de factori. Plecând de la regula de adunare a varianțelor, se pot calcula doi coeficienți:

– coeficientul influenței factorului de grupare (k_1), calculat după relația:

$$k_1 = \frac{\delta^2}{\sigma^2} \cdot 100;$$

– coeficientul influenței factorilor întâmplători (k_2), calculat după relația:

$$k_2 = \frac{\overline{\sigma^2}}{\sigma^2} \cdot 100.$$

Suma celor doi coeficienți este egală cu *1* sau *100%*:

$$(k_1 + k_2 = 100\%).$$

Observație! Cu cât valoarea lui $\left(\frac{\delta^2}{\sigma^2} \cdot 100\right) > \left(\frac{\overline{\sigma^2}}{\sigma^2} \cdot 100\right)$, cu atât factorul de grupare are o influență mai mare asupra variației caracteristicii de distribuție a colectivității.

Indicatorii factoriali ai dispersiei în cazul unei variabile alternative (dichotomice)

Varianța de grupă se calculează după relația:

$$\sigma_p^2 = p_j \cdot q_j.$$

Media varianțelor de grupă ($\overline{\sigma_p^2}$) se calculează după relația:

$$\overline{\sigma_p^2} = \frac{\sum_{j=1}^{k} p_j q_j \cdot n_{\bullet j}}{\sum_{j=1}^{k} n_{\bullet j}}, \ j = \overline{1,k} \text{ grupe}.$$

Varianța mediilor de grupă față de media generală (δ_p^2) se calculează după relația:

$$\delta_p^2 = \frac{\sum_{j=1}^{k} (p_j - p)^2 n_{\bullet j}}{\sum_j n_{\bullet j}}, \text{ unde } p = \frac{\sum_{j=1}^{k} p_j n_{\bullet j}}{\sum_{j=1}^{k} n_{\bullet j}}.$$

Relația de adunare a varianțelor este:

$$\sigma_p^2 = \overline{\sigma_p^2} + \delta_p^2 .$$

Exemplu de obținere a indicatorilor factoriali ai dispersiei, folosind statisticile calculate cu ajutorul SPSS. Considerăm output-ul privind venitul lunar pe categorii de vârstă, prezentat în figura 5.18.

Se cere:

1. Să se măsoare dispersia sub influența factorilor întâmplători pe ansamblul datelor;
2. Să se determine dispersia sub influența factorilor esențiali (de grupare);
3. Să se măsoare dispersia pe ansamblu, folosind regula de adunare a varianțelor.

Rezolvare

1. Influența factorilor întâmplători asupra variației unei caracteristici se măsoară prin varianța de grupă și media varianțelor de grupă.

Influența factorilor întâmplători pe total se măsoară prin *media varianțelor de grupă* ($\overline{\sigma^2}$):

$$\overline{\sigma^2} = \frac{\sum_j \sigma_j^2 n_{.j}}{\sum_j n_{.j}} = \frac{4{,}323 \cdot 112 + 7{,}044 \cdot 266 + 0{,}909 \cdot 22}{400} = \frac{2377{,}878}{400} = 5{,}944695$$

$$\overline{\sigma} = \sqrt{\overline{\sigma^2}} = \sqrt{5{,}944695} = 2{,}43817 \text{ milioane lei}$$

Intervalul mediu de variație sub influența factorilor întâmplători (alții decât apartenența la grupa de vârstă) este, pe total:

$$\bar{x} \pm \bar{\sigma} = \begin{cases} 2,87 - 2,44 = 0,43 \\ 2,87 + 2,44 = 4,31 \end{cases} \text{ milioane lei.}$$

2. *Dispersia sub influența factorului de grupare* este exprimată prin *varianța mediilor de grupă față de media generală* (δ^2):

$$\delta^2 = \frac{\sum_j (\bar{x}_j - \bar{x})^2 \cdot n_{\bullet j}}{\sum_j n_{\bullet j}} =$$

$$= \frac{(1,96 - 2,87)^2 \cdot 112 + (3,35 - 2,87)^2 \cdot 266 + (1,64 - 2,87)^2 \cdot 22}{400} = 0,4682935$$

$$\delta = \sqrt{\delta^2} = \sqrt{0,4682935} = 0,6843 \text{ milioane lei.}$$

Intervalul mediu de variație, sub influența factorilor esențiali (apartenența la grupa de vârstă), este:

$$\bar{x} \pm \delta = \begin{cases} 2,87 - 0,68 = 2,19 \\ 2,87 + 0,68 = 3,55 \end{cases} \text{ milioane lei.}$$

3. *Dispersia sub influența ambelor categorii de factori* se măsoară prin *varianța totală* (σ^2).

Folosind *regula de adunare a varianțelor*:

$$\sigma^2 = \overline{\sigma^2} + \delta^2 = 5,94 + 0,468 = 6,41.$$

Observație! Diferența de valoare dintre varianța totală a venitului lunar, prezentată în output (6,399), și varianța rezultată din regula de adunare a componentelor sale (6,41) este datorată aproximărilor de calcul:

$$\sigma = \sqrt{\sigma^2} = 2,53 \text{ milioane lei.}$$

Intervalul mediu de variație, sub influența atât a apartenenței la grupa de vârstă, cât și a factorilor întâmplători care au acționat asupra distribuției în cadrul fiecărei grupe, este:

$$\bar{x} \pm \sigma = \begin{cases} 2,87 - 2,53 = 0,34 \\ 2,87 + 2,53 = 5,4 \end{cases} \text{ milioane lei.}$$

CAPITOLUL 6

DISTRIBUȚIA NORMALĂ

- **Distribuția normală**
- **Distribuția normală standard**
- **Calculul probabilităților pentru distribuții normale folosind SPSS**
- **Verificarea normalității unei distribuții folosind SPSS**

Distribuția normală este cea mai cunoscută și mai folosită distribuție statistică, și aceasta din cel puțin două motive. Primul – foarte multe variabile statistice, cum ar fi greutatea, înălțimea, vârsta oamenilor, sau numeroase variabile specifice lumii afacerilor, de exemplu, venitul populației, profitul firmelor, urmează o distribuție normală; al doilea – câteva statistici importante, cum ar fi media de selecție, se distribuie după un model normal. Distribuția normală se constituie ca bază pentru statistica inferențială clasică, folosirea rezultatelor cercetărilor prin sondaj plecând de la ipoteza că eșantioanele observate provin din populații distribuite normal.

În acest capitol vom prezenta caracteristicile prin care poate fi identificată o distribuție normală, vom exemplifica cum se calculează, manual și în SPSS, probabilitățile pentru distribuții normale, vom vedea cum este folosită distribuția normală pentru aproximarea altor distribuții de probabilitate.

6.1 Distribuția normală

Simbolizare.

Pentru o variabilă X, care urmează o lege normală (sau legea Gauss-Laplace), de parametri μ și σ^2, vom folosi notația: $X \sim N(\mu,\sigma^2)$.

6.1.1 Funcția de densitate de probabilitate și funcția de repartiție

O variabilă aleatorie X este distribuită după o lege normală generalizată dacă are o *funcție de densitate de probabilitate* de forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ x \in R,$$

unde:

e – constantă matematică aproximată prin 2,71828;

π – constantă matematică aproximată prin 3,14159;

μ – media populației;

σ – abaterea medie pătratică (deviația standard);

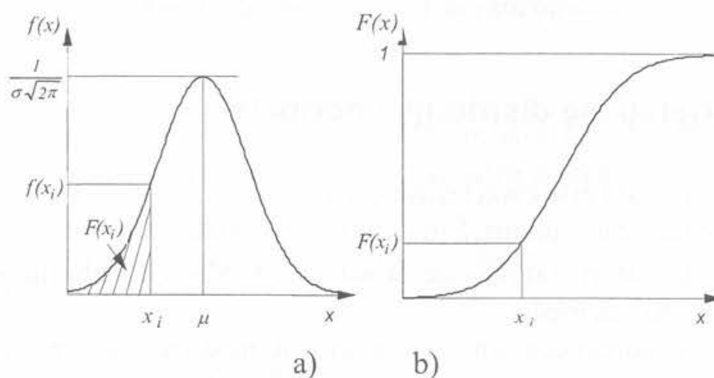
x – orice valoare a variabilei continue $X(-\infty < X < \infty)$.

Observație! De regulă, notația folosită pentru o variabilă este o literă majusculă, de exemplu X, iar pentru o valoare a variabilei se folosește o literă minusculă, de exemplu x_i. În SPSS, pentru x_i se folosește notația q.

Funcția de densitate de probabilitate este reprezentată grafic prin *curba densității de probabilitate*, curbă cu alură de clopot (vezi figura 6.1 a).

Funcția de repartiție a legii normale generalizate are forma prezentată în figura 6.1 b și este definită prin relația:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt .$$



a) b)

Figura 6.1 *Curba normală*: *a)* Densitatea de probabilitate *și b)* Funcția de repartiție

Aria de sub curba densității de probabilitate este egală cu unu.

Pentru o variabilă continuă, se poate calcula probabilitatea ca o valoare să fie cuprinsă într-un interval. Probabilitatea ca o variabilă aleatorie continuă să ia o valoare exactă este egală cu zero.

În modelul $f(x)$, e și π sunt constante, prin urmare o distribuție normală este complet descrisă numai de medie și abaterea medie pătratică (numită în SPSS deviație standard). Folosind diferite combinații de medie și deviație standard se pot genera diferite distribuții de probabilitate normale.

De exemplu, pentru aceeași medie, dar cu două deviații standard diferite, se obțin două distribuții diferite (vezi figura 6.2)

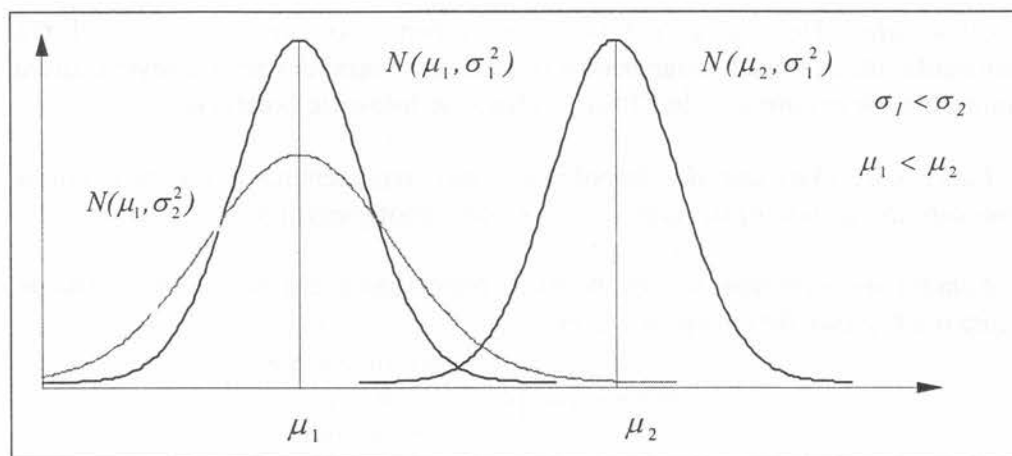


Figura 6.2. *Curbe normale cu aceeaşi medie şi cu deviaţii standard diferite, respectiv cu medii diferite şi aceeaşi deviaţie standard*

6.1.2 Proprietăţi ale distribuţiei normale

O distribuţie normală este caracterizată prin:

- Curba normală este simetrică în raport cu ordonata valorii $x = \mu$, 50% din observaţii au valori mai mici decât valoarea medie a distribuţiei şi 50% au valori mai mari ca media;
- Indicatorii tendinţei centrale (media, modul, mediana) au aceeaşi valoare;
- Intervalul interquartilic este cuprins între două treimi din deviaţia standard sub medie şi două treimi din deviaţia standard peste medie;
- Variabila aleatoare într-o distrubuţie normală are o amplitudine infinită ($-\infty < X < \infty$), curba normală nu atinge niciodată axa orizontală a graficului (abscisa). Când $x \to \pm\infty$, funcţia *f(x)* tinde spre zero (se apropie asimptotic de axa *Ox*);
- Funcţia f(x) este maximă pentru $x = \mu$ şi se diminuează pe măsură ce valorile variabilei se depărtează de medie;
- Curba densităţii de probabilitate f(x) are puncte de inflexiune când $x = \mu \pm \sigma$; o distribuţie normală este unic determinată de medie şi de varianţă: $X \sim N(\mu; \sigma^2)$.

În practica economică, numeroase variabile au distribuţii care aproximează proprietăţile unei distribuţii normale teoretice, prezentând anumite grade de asimetrie şi boltire, o combinaţie infinită a parametrilor (medie şi abaterea medie pătratică) şi au o amplitudine finită în raport cu fenomenul observat

(de regulă, o amplitudine egală cu intervalul media plus/minus de 3 ori deviaţia standard).

6.2 Distribuţia normală standard

Distribuţia normală standard este distribuţia variabilei normale centrate reduse Z, numită variabilă aleatorie standard. Valorile variabilei Z, numită şi *variabilă scor*, se obţin ca diferenţă dintre valorile unei variabile X şi media populaţiei divizată prin deviaţia standard σ, respectiv media eşantionului şi abaterea medie pătratică corespunzătoare, după relaţiile:

$$Z = \frac{X - \mu}{\sigma}, \text{ respectiv } Z = \frac{X - \bar{x}}{s}.$$

6.2.1 Funcţia de densitate de probabilitate a distribuţiei normale standard şi funcţia de repartiţie a acesteia

Funcţia de densitate de probabilitate a distribuţiei normale standard *f(z)* şi funcţia de repartiţie *F(z)* sunt definite de relaţiile:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2} \; ; \quad F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt.$$

Proprietăţi:

$$f(z) > 0;$$
$$f(-z) = f(z);$$
$$\lim_{z \to \pm\infty} f(z) = 0;$$
$$\int_{-\infty}^{\infty} f(z)dz = 1.$$

Parametrii distribuţiei sunt:

$$-\text{media } M(Z) = \int_{-\infty}^{+\infty} z \cdot f(z)dz = 0;$$

– varianța $V(Z) = \int_{-\infty}^{+\infty} z^2 \cdot f(z)dz = 1.$

O distribuție normală standard este o distribuție a cărei variabilă Z are întotdeauna media $\mu = 0$ și deviația standard $\sigma = 1$, adică urmează o lege de distribuție normală cu media egală cu zero și deviația standard egală cu unu: $Z \sim N(0;1)$.

Observație! În practica statistică, este important să știm cum se folosește variabila Z și tabelele distribuției normale, precum și conversia din scoruri Z în percentile și invers.

6.2.2 Standardizarea unei variabile X

Standardizarea unei variabile (transformarea unei distribuții normale într-o distribuție normală standard) presupune trecerea de la o distribuție normală $X \sim N(\mu, \sigma^2)$ la o distribuția standard $Z \sim N(0;1)$, adică efectuarea unei transformări asupra tuturor valorilor unei distribuției, după relația:

$$z_i = \frac{x_i - \bar{x}}{\sigma},$$

unde:

x_i sunt valori ale seriei observate;

$\bar{x}$ și σ, valoarea medie și abaterea standard ale seriei observate.

Dacă Z este o variabilă normală standard, atunci variabila $X = \mu + \sigma \cdot Z$ urmează o lege de distribuție normală generalizată, definită de următorii parametri:

$$M(X) = M(\mu + \sigma \cdot Z) = M(\mu) + \sigma \cdot M(Z) = \mu;$$

$$V(X) = V(\mu + \sigma \cdot Z) = V(\mu) + \sigma^2 \cdot V(Z) = \sigma^2.$$

Deoarece media variabilei Z este egală cu zero, valoarea $z = 0$ corespunde cu valoarea medie a seriei de origine (dacă $z = 0$, $x_i - \bar{x} = 0$), deci $x_i = \bar{x}$ pentru $z = 0$ (vezi figura 6.3).

Oricare valoare x_i superioară mediei are o valoare corespunzătoare în Z superioară lui 0, adică pozitivă, respectiv orice valoare x_i inferioară mediei are o valoare corespunzătoare în Z mai mică decât 0, adică negativă.

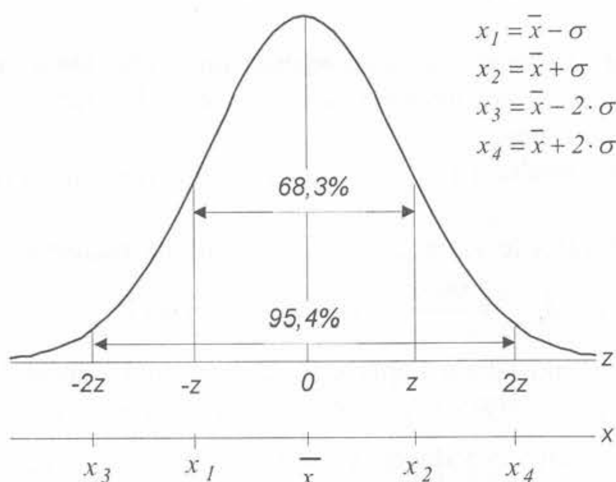


Figura 6.3 *Corespondența dintre o distribuție Z și o distribuție a unei variabile normale X*

Interpretarea unei distribuții normale este facilitată prin transformarea sa într-o distribuție Z, prin faptul că pentru variabila normală standard Z s-au construit tabele, cu ajutorul cărora se pot citi probabilitățile corespunzătoare valorilor z_i (vezi tabelele Laplace).

Pentru a obține probabilitatea unei distribuții normale, este necesar ca valorile variabilei X să fie exprimate în unități de abateri standard față de medie, adică variabila X să se standardizeze:

$$X = \mu + Z \cdot \sigma .$$

De exemplu, pentru $Z=1$, dacă $X \sim N(\mu, \sigma^2)$, se poate afla probabilitatea ca X să ia valori în intervalul definit de valoarea medie, plus/minus o deviație standard.

Dacă $X = \mu - \sigma$, atunci $Z = \frac{(\mu - \sigma) - \mu}{\sigma} = -1$. Similar, dacă $X = \mu + \sigma$, atunci $Z = 1$.

Astfel,

$$P(\mu - \sigma \le X \le \mu + \sigma) = P(-1 \le Z \le 1) = 2P(Z \le 1) - 1 = 0,6826 \approx 2/3 .$$

Pentru Z = *2*, respectiv Z = *3*, aflăm probabilitatea în mod analog, şi anume:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2) = 0{,}954,$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-3 \leq Z \leq 3) = 0{,}998.$$

În mod practic, pentru *transformarea unei distribuţii statistice într-o distribuţie teoretică* şi interpretarea acesteia în termeni probabilistici parcurgem următorii paşi:

1. Se calculează media ($\bar{x}$) şi abaterea medie pătratică (σ) pentru seria observată;
2. Se calculează valorile variabilei normale centrate reduse Z corespunzătoare valorilor x_i, adică $z_i = \frac{(x_i - \bar{x})}{\sigma}$, numite şi *scoruri Z*;
3. Se citeşte, folosind tabela Laplace-Gauss[1], numită Tabela 1, probabilitatea corespunzătoare: $P(Z < z_i)$, adică probabilitatea ca o unitate din colectivitate să aibă o valoare *X* inferioară valorii x_i considerată. Tabela indică valoarea suprafeţei cuprinsă între curbă, axa *Ox* şi ordonatele în $x = 0$ şi $x_i = z_i$ (vezi figura 6.4).

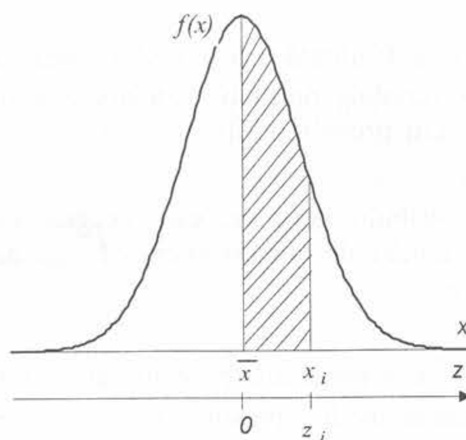


Figura 6.4
Probabilitatea $P(0 < Z < z_i)$

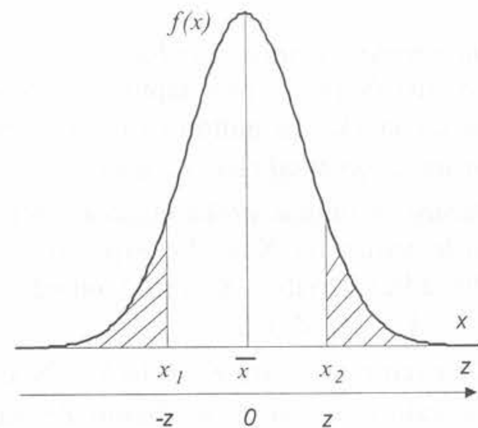


Figura 6.5
Probabilitatea $P(Z > z_i) + P(Z < -z_i)$

1. Tabelele furnizează direct datele reprezentând suprafaţa cuprinsă între curba densităţii de probabilitate (numită „curba Gauss"), axa X şi două ordonate considerate, suprafaţă care matematic se poate afla cu ajutorul calculului integral.

Se poate folosi o a doua tabelă Gauss care indică valori corespunzătoare ariei exterioare dreptelor $x_1 = -z,\ x_2 = z$, arătând probabilitatea $P(Z > z) + P(Z < -z)$ care este egală cu $1 - 2 \cdot P(Z < z)$ (vezi figura 6.5).

6.2.3. Obținerea valorilor variabilei Z folosind SPSS

Valorile variabilei Z (scorurile z) se pot obține, folosind SPSS, parcurgând următorii pași:

- Se selectează succesiv meniul *Analyze* → comanda *Descriptive Statistics* → opțiunea *Descriptives*;
- Din fereastra deschisă (*Descriptives*) se alege variabila pe care dorim să o standardizăm (să o transformăm în scoruri z) și o mutăm în zona *Variable(s)*;
- Bifăm caseta de validare *Save Standardized Values as Variables*;
- Activăm butonul de comandă *OK*.

Variabila standardizată este salvată în *Data File* (în exemplul nostru *Tapestry.sav*), în partea dreaptă a foii *Data View*, și este automat numită *z*, urmat de numele variabilei (în exemplul nostru *zvlunar* – vezi figura 6.6).

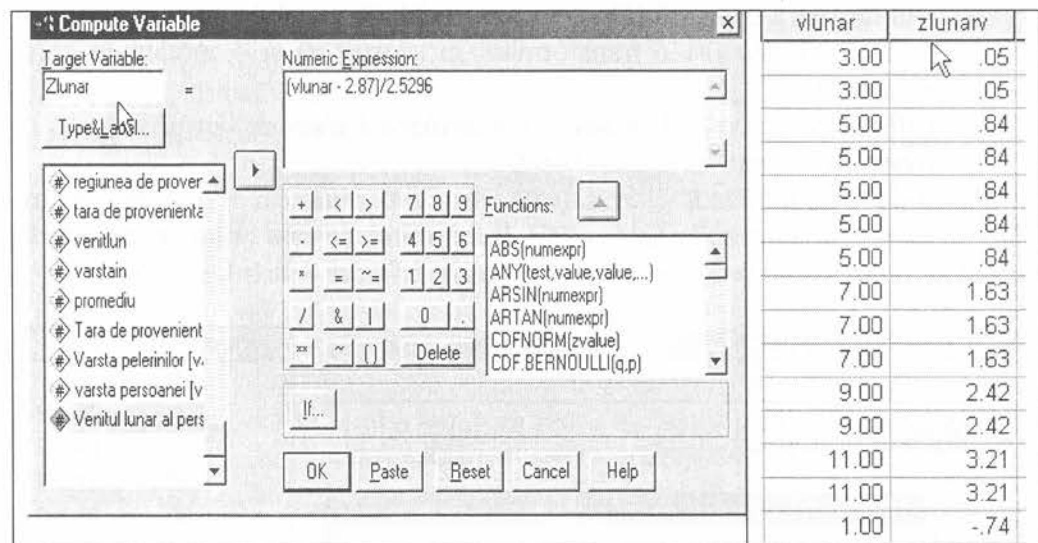
vlunar	zvlunar
3.00	.05139
3.00	.05139
5.00	.84202
5.00	.84202
5.00	.84202
5.00	.84202
5.00	.84202
7.00	1.63264
7.00	1.63264
7.00	1.63264
9.00	2.42327
9.00	2.42327
11.00	3.21389
11.00	3.21389
1.00	-.73923

Figura 6.6. *Obținerea valorilor standardizate ale unei variabile X prin: meniul* Analyze → *comada* Descriptive Statistics → *opțiunea* Descriptives

Valorile standardizate se mai pot obţine, plecând de la media populaţiei şi deviaţia standard, *folosind formula pentru scorul z*.

Demersul de urmat este:

- Meniul *Transform* → comanda *Compute*;
- Scriem numele noii variabile (*zvlunar*) în zona *Target Variable*;
- Dublu clic pe variabila pe care dorim să o standardizăm pentru a o introduce în zona *Numeric Expression*;
- Scriem *formula* scorului *z* pentru această variabilă. Astfel, pentru o medie a populaţiei egală cu 2,87 şi o abaterea standard egală cu 2,5296, scriem formula: *(vlunar – 2,87)/2,5296* în zona *Numeric Expression* (vezi figura 6.7).



vlunar	zlunarv
3.00	.05
3.00	.05
5.00	.84
5.00	.84
5.00	.84
5.00	.84
5.00	.84
7.00	1.63
7.00	1.63
7.00	1.63
9.00	2.42
9.00	2.42
11.00	3.21
11.00	3.21
1.00	-.74

Figura 6.7. *Obţinerea valorilor standardizate ale unei variabile X prin: meniul* Transform → *comanda* Compute

Observaţie! Standardizarea realizată prin cele două căi duce la obţinerea aceloraşi rezultate (vezi valorile standardizate *zvlunar* şi *zlunarv*, prezentate în figurile 6.6 şi 6.7).

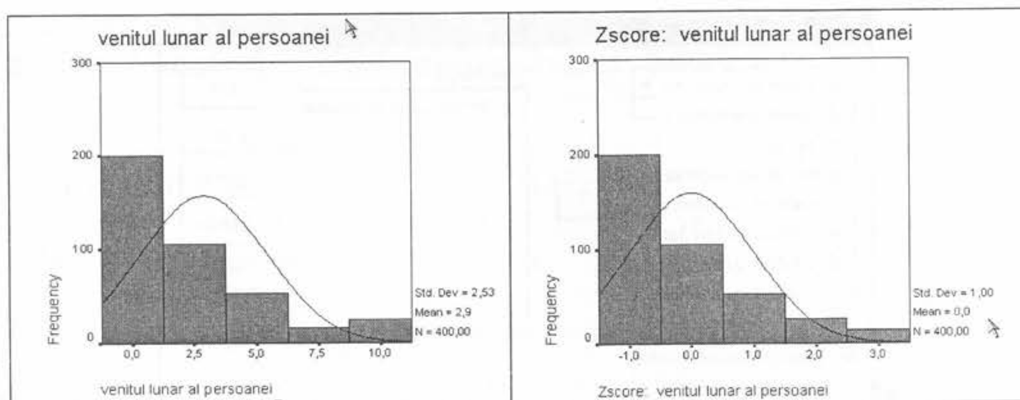


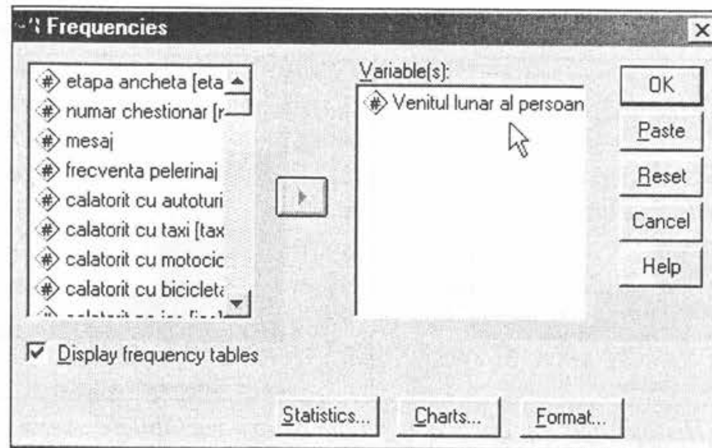
Figura 6.8 *Histogramele şi curbele normale pentru variabilele „venitul lunar" şi „Zscor" corespunzător*

Observaţie! Variabila Z are media egală cu zero şi abaterea standard egală cu unu. Se observă că, pentru o variabilă X, în cazul dat venitul lunar, şi variabila Z corespunzătoare, se obţin histograme şi curbe ale frecvenţelor cu aceeaşi alură; mediile şi abaterile standard ale celor două distribuţii coincid (vezi figura 6.8).

6.3 Calculul probabilităţilor pentru distribuţii normale folosind SPSS

6.3.1 Aproximarea probabilităţii pentru o variabilă aleatorie normală pe baza frecvenţelor relative cumulate

Pe baza frecvenţelor relative (proporţii) cumulate se poate afla probabilitatea ca o valoare a unei variabile aleatorie distribuită normal să aparţină unui interval, efectuând în SPSS următorul demers: meniul *Analyze* → comanda *Descriptive Statistics* → opţiunea *Frequencies* (vezi figura 6.9).



Venitul lunar al persoanei

Venit	Frequency	Percent	Valid Percent	Cumulative Percent
0 - 2	200	50.0	50.0	50.0
2 - 4	106	26.5	26.5	76.5
4 - 6	53	13.3	13.3	89.8
6 - 8	16	4.0	4.0	93.8
8 - 10	11	2.8	2.8	96.5
10 şi peste	14	3.5	3.5	100.0
Total	400	100.0	100.0	

Figura 6.9 *Aflarea procentelor cumulate în SPSS – Output-ul* Frequencies

De exemplu, folosind baza de date *tapesty.sav*, care este procentul persoanelor anchetate care au un venit lunar < 4 milioane lei? Dar al celor care au un venit lunar în intervalul 8-10 milioane lei?

Procentul observaţiilor totale corespunde ariei de sub curba asociată distribuţiei normale.

Din output-ul prezentat în figura 6.9, se poate afla că din cele 400 de persoane anchetate, 76,5% au venitul lunar < 4 milioane lei.

Procentul observaţiilor corespunzător unui interval se află prin scăderea procentelor cumulate corespunzătoare celor două limite ale intervalului dorit; astfel, pentru intervalul 8-10 milioane, aflăm 96,5-93,8 = 2,7%. Procentele cumulate corespund probabilităţii ca o valoare să se găsească într-un anumit interval, de exemplu: P(8 < X < 10) = [P(X < 10 milioane = 0,965) – P(X < 8 = 0,938) = 0,027].

6.3.2 Calculul probabilităţilor pentru o variabilă aleatorie normală folosind funcţiile disponibile în SPSS

Funcţiile disponibile în SPSS folosite în calculul probabilităţilor pentru o variabilă normală sunt: CDF.NORMAL, CDFNORM, PROBIT şi IDF.NORMAL. Accesul la aceste funcţii, în SPSS, se face urmând demersul:

Meniul *Transform* → comanda *Compute.*

O distribuţie normală este unic determinată de medie (μ) şi de varianţă (σ^2); astfel, $X \sim N(\mu;\sigma^2)$. Cunoaşterea valorilor μ şi σ^2 permite să se determine probabilitatea pe care o are variabila aleatorie de a aparţine unui interval oarecare:

$P(X < a)$ sau $P(X > a)$ sau $P(a < X < b) = P(X<b) - P(X < a)$,

unde: a şi b sunt numere.

În SPSS, calculul acestor probabilităţi se poate face direct folosind funcţia CDF.NORMAL, fără a mai fi necesară standardizarea variabilei X (transformarea acesteia în scoruri Z) înainte de calculul probabilităţii. Sintaxa acestei funcţii este CDF.NORMAL (q, μ, σ) (vezi figura 6.10).

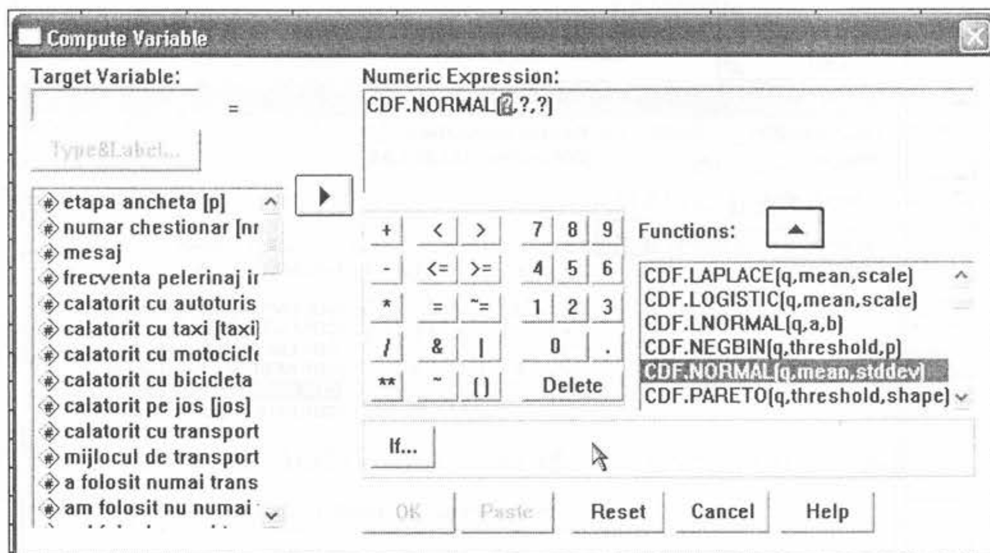


Figura 6.10. *Alegerea funcţiilor în Fereastra dialog* Compute Variable

Calculul probabilității P(X < a)

Exemplificăm calculul probabilității *P(X < a)*, unde *X* este o variabilă aleatorie normală, cu media μ și abaterea standard σ, adică probabilitatea ca *X* să fie mai mică decât un număr *a*; folosim funcția CDF.NORMAL din SPSS, considerând variabila „Venitul lunar" din *tapestry. sav*, cu *valoarea medie* de *2,87* milioane lei și *abaterea standard* de *2,5296* milioane lei.

Pașii de urmat sunt:

- Se deschide fereastra *Data Editor*, în care se introduce o valoare a variabilei în prima celulă din foaia de lucru;
- Se alege meniul *Transform* → comanda *Compute*;
- În zona *Target Variable* din fereastra *Compute Variable* introducem numele variabilei pentru a cărei valoare dorim să calculăm probabilitatea, de exemplu, „prob_vl" (vezi figura 6.11);
- În zona *Numeric Expression* introducem expresia funcției, selectată din lista *Functions*, CDF.NORMAL(q, mean, stddev), unde *q* este o valoare *a* a variabilei X. Pentru exemplul dat, CDF.NORMAL(4, 2.87, 2.53);
- Prin butonul *OK*, se comandă calculul propriu-zis al probabilității.

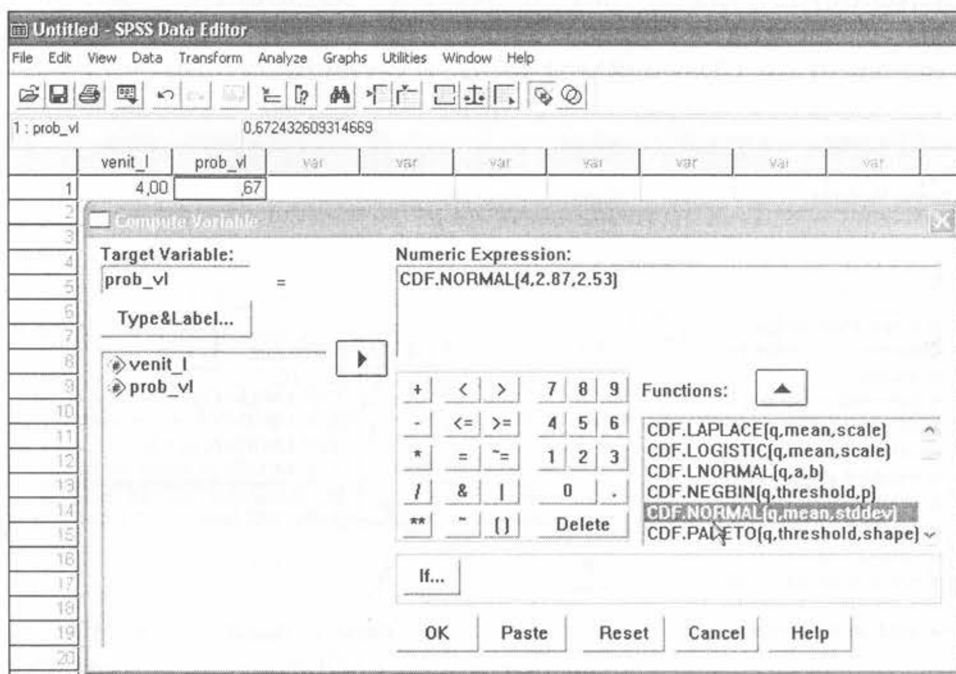


Figura 6.11. *Calculul probabilității cu ajutorul funcției* CDF.NORMAL

Valoarea P(X < 4) = 0,67, calculată pentru variabila $X \sim N(\mu, \sigma^2)$, respectiv $X \sim N(2.87, 2.53^2)$, apare în celula de sub numele variabilei „prob_vl" din foaia de lucru a ferestrei *Data Editor* (vezi figura 6.11). Când valoarea probabilității este selectată, în celula de editare este afișată valoarea probabilității cu 15 zecimale.

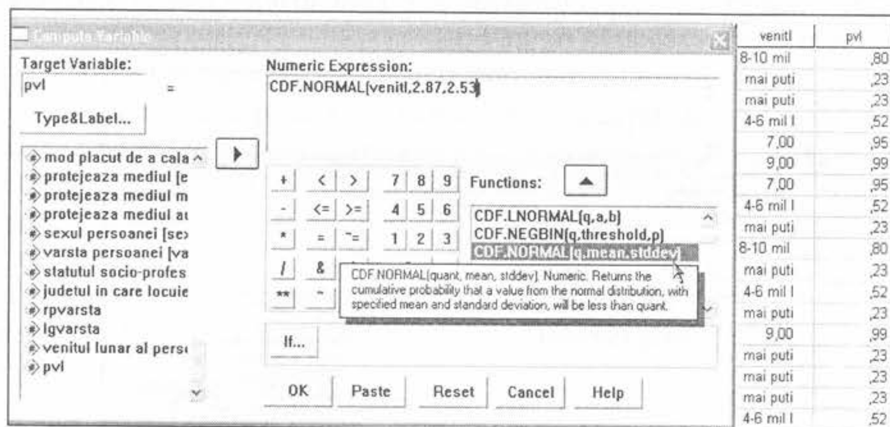


Figura 6.12 *Aflarea probabilității P(X < a) pentru valorile venitului lunar folosind funcția* CDF.NORMAL *(q, mean, stddev)*

Probabilitățile pentru valorile variabilei *vlunar* sunt prezentate în figura 6.12 și se află urmând același demers.

Calculul probabilității P(X > a) și al probabilității P(a < X < b)

Calculul probabilității *P(X > a)* presupune să se găsească 1 – *P(X < a)*. Realizarea acestei operații cere același demers prezent mai sus, cu deosebirea că în zona *Numeric Expression* introducem: 1 – CDF.NORMAL(a, μ, σ).

Aflarea probabilității *P(a < X < b)* se bazează pe relația:

$P(a < X < b) = P(X < b) - P(X < a)$

și presupune calculul probabilităților *P(X < a)* și *P(X < b)*, după demersul prezentat anterior.

6.3.3 Calculul probabilităților pentru o variabilă normală standard (Z)

O variabilă normală standard Z este o variabilă cu media zero și abaterea standard 1, $Z \sim N(0;1)$.

Calculul probabilităților pentru o astfel de variabilă, în SPSS, presupune folosirea funcției CDFNORM. Sintaxa acestei funcții este CDFNORM (q) (vezi figura 6.13) unde q este o valoare „a" a variabilei Z, pentru care se calculează P(Z < a).

Exemplificăm calculul probabilității corespunzătoare variabilei Z mai mică decât 2,5, adică P(Z < 2,5). Demersul este asemănător celui folosit pentru o variabilă X și este prezentat în figura 6.13.

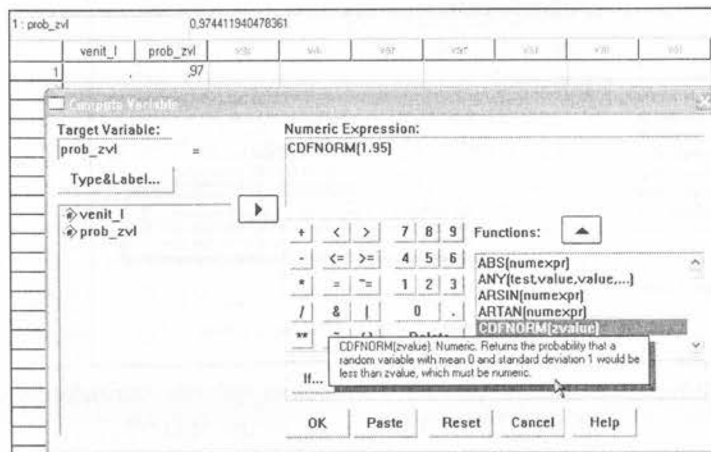


Figura 6.13 *Aflarea probabilității* P(Z < a) *folosind funcția* CDFNORM

6.3.4 Aflarea valorilor variabilei Z și a valorilor unei variabile normale X pentru probabilități cunoscute

Cazul variabilei normale standard Z

În SPSS, pentru calculul valorilor variabilei normale standard Z, se folosește funcția PROBIT. Această funcție are sintaxa PROBIT(prob) și dă valoarea z_0 a variabilei Z a cărei probabilitate este egală cu *prob*, adică se calculează z_0, astfel ca $P(Z < z_0)$ = prob.

De exemplu, pentru a afla $P(Z < z_0)$ =0,95, introducem în *Numeric Expression* din fereastra *Compute Variable* expresia *PROBIT (0,95)*. Se obține astfel pentru z_0 o valoare egală cu 1,64.

Pentru $P(Z > z_0) = prob$, valoarea z_0 se află folosind sintaxa PROBIT(1 – prob).

Cazul unei variabile normale X

În cazul unei variabile normale X, calculul unei valori „a" a variabilei pentru o probabilitate cunoscută, adică $P(X < a) = prob$, se efectuează, în SPSS,

folosind funcția IDF.NORMAL, a cărei sintaxă este IDF.NORMAL(prob, μ,σ), unde μ,σ reprezintă media și abaterea standard.

Pentru $P(X > a) = prob$, se folosește IDF.NORMAL(1 – prob, μ,σ).

6.4 Verificarea normalității unei distribuții folosind SPSS

Majoritatea testelor parametrice cer îndeplinirea condiției de normalitate pentru variabilele considerate, ipoteza de normalitate a unei distribuții fiind una dintre ipotezele comune care se presupun în procesul de inferență statistică. Modelarea statistică cere verificarea normalității variabilelor implicate. Fără respectarea acestei ipoteze, nu ar fi valide interpretarea și inferența bazate pe astfel de modele.

Prin urmare, este deosebit de important ca, înainte de efectuarea procesului de inferență, să se determine dacă eșantionul observat provine dintr-o populație normal distribuită.

În SPSS, se pot folosi două căi de verificare a normalității unei distribuții, și anume:

- vizualizarea grafică a diferențelor dintre o distribuție empirică și distribuția teoretică, folosind *histograma*, *boxplot*, *P-P plot* și *Q-Q plot*;
- aprecierea numerică a abaterilor distribuției empirice de la distribuția teoretică, folosind indicatori din *statistica descriptivă* și *teste statistice*.

6.4.1. Procedeul histogramei

Folosirea histogramei pentru a diagnostica dacă o distribuție este normală presupune *compararea histogramei variabilei observate* cu *modelul curba Gauss*.

Obținerea acestor diagrame în SPSS presupune următorul demers: meniul *Graphs* → comanda *Histogram*.

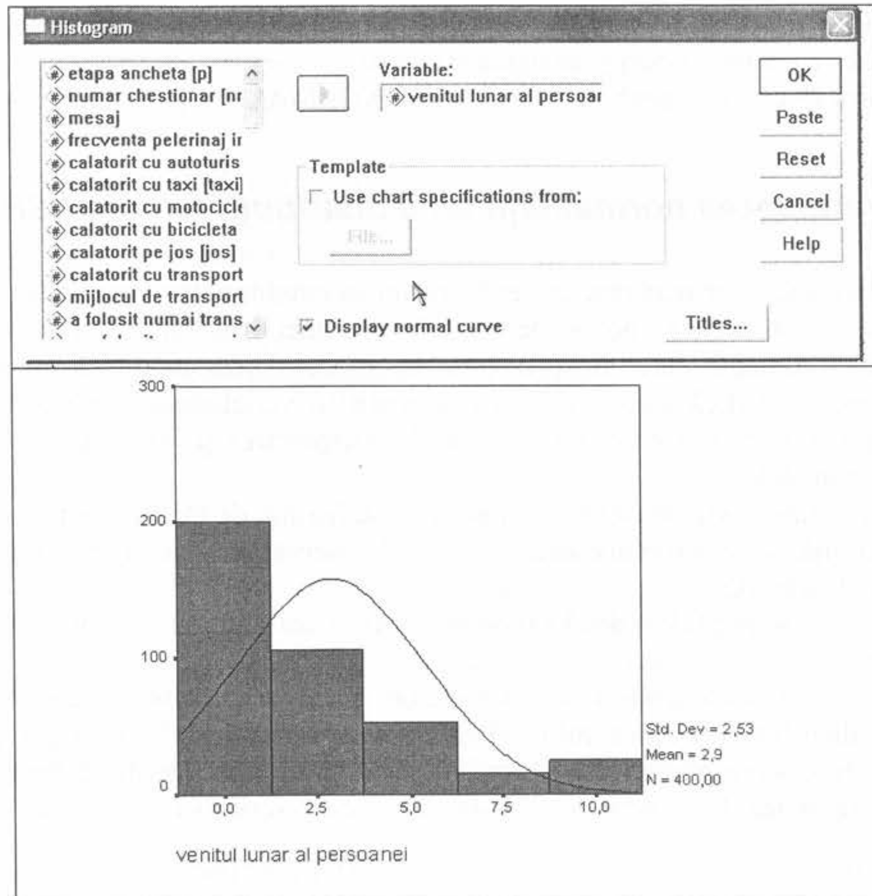


Figura 6.14 *Obținerea histogramei și curbei normale prin demersul: Meniul* Graphs → *comanda* Histogram

În fereastra *Histogram* se bifează caseta de validare *Display normal curve* și se activează butonul de comandă *OK* pentru a obține output-ul dorit (vezi figura 6.14). Ca urmare a operației de bifare în caseta de validare *Display normal curve*, se adaugă o curbă normală la histogramă, cu aceeași medie și aceeași varianță corespunzătoare distribuției empirice.

De asemenea, poate fi folosită procedura *Frequencies*, urmând demersul:

- Se selectează succesiv meniul *Analyze* → comanda *Descriptive Statistics* → opțiunea *Frequencies*;
- Din fereastra *Frequencies*, după selectarea variabilei/variabilelor, se activează butonul de comandă *Charts* care deschide fereastra *Frequencies: Charts*;

- În fereastra *Frequencies: Charts* bifăm caseta de validare *With normal curve* și acționăm butonul de comandă *Continue* pentru a reveni la fereastra *Frequencies*;
- Se activează butonul *OK*, care comandă SPSS obținerea output-ului dorit (vezi figura 6.15).

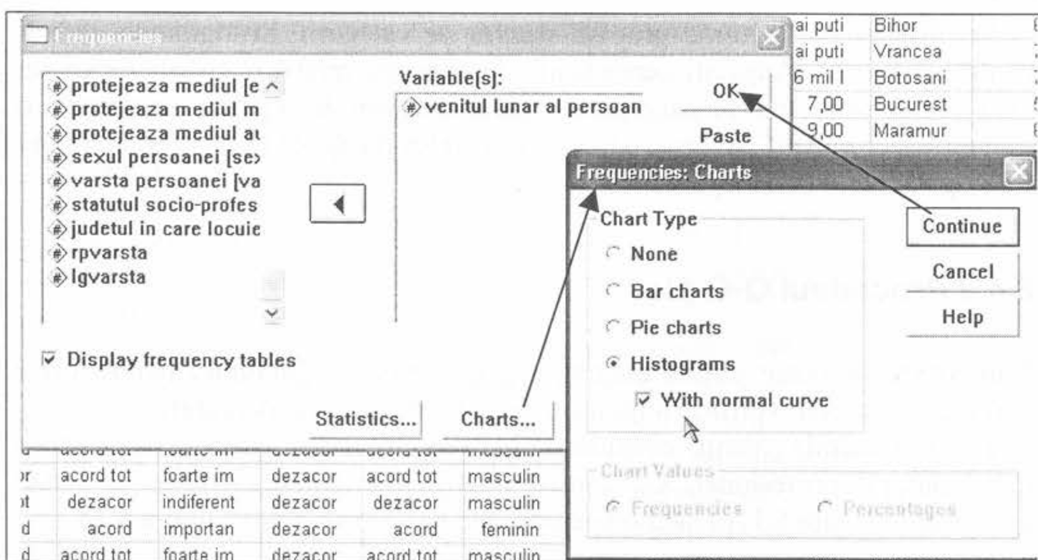


Figura 6.15 *Obținerea histogramei și curbei normale prin demersul: meniul* Analyse → *comanda* Descriptive Statistics → *opțiunea* Frequencies

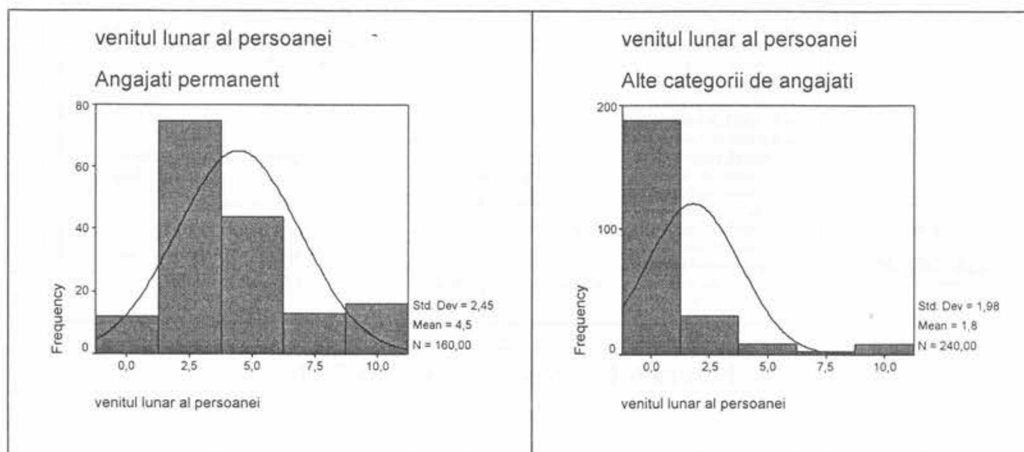


Figura 6.16 *Histograme pentru venitul lunar, pe categorii de angajați*

Histograma din figura 6.16 relevă o distribuţie cu un grad mare de asimetrie; normalitatea distribuţiei poate fi pusă sub semnul întrebării. În astfel de situaţii, este posibil să se grupeze datele în funcţie de un factor determinant (cum ar fi în cazul nostru „statutul profesional"), folosind funcţia *Split File* din meniul *Data*, şi să se construiască histograme pentru fiecare categorie (vezi figura 6.16).

Se observă că, în urma grupării datelor pe categorii, histogramele diferă. Pentru categoria „Angajaţi permanent", histograma relevă o distribuţie cu un grad mic de asimetrie, pe când pentru „Alte categorii de angajaţi" se observă o asimetrie accentuată. Aceeaşi situaţie este relevată de *Q-Q plot* (vezi figura 6.18) şi *P-P plot* (vezi figura 6.19).

6.4.2 Procedeul Q-Q plot

Prin SPSS, se poate obţine diagrama *Q-Q (Quantile Quantile)* pentru orice variabilă, în scopul verificării ipotezei de normalitate (vezi paragraful 4.2.2).

Pentru variabila „Venit" considerată în exemplul nostru, grupată pe categorii după statutul profesional, s-a construit *Q-Q plot* parcurgându-se demersul prezentat în figura 6.17. Diagramele obţinute sunt prezentate în figura 6.18.

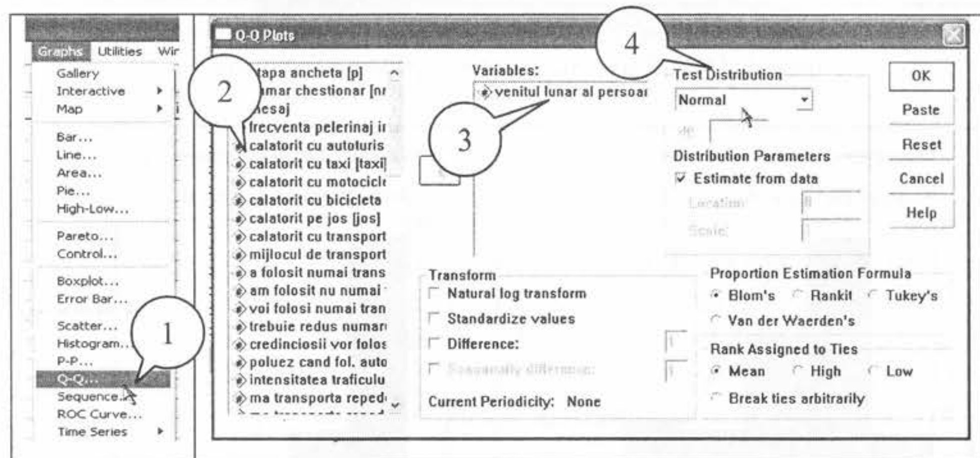


Figura 6.17 *Fereastra dialog* Q-Q plot

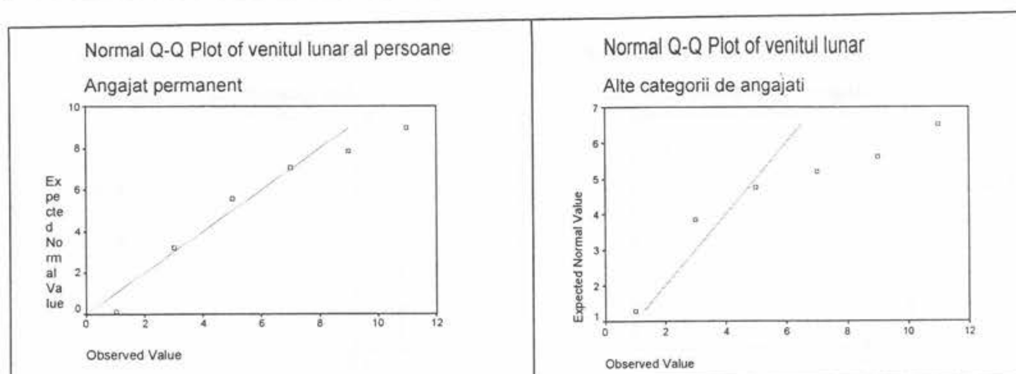


Figura 6.18 Q-Q plot *pentru variabila „Venit", pe categorii de angajaţi*

Q-Q plot compară valorile ordonate ale variabilei observate cu valorile quantilice ale distribuţiei teoretice specificate (în cazul nostru, distribuţia normală).

Dacă distribuţia variabilei testate este normală, atunci punctele Q-Q conturează o linie care se suprapune cu dreapta care reprezintă distribuţia teoretică, adică trece prin origine şi are panta egală cu unu.

Pentru exemplul dat, Q-Q plot arată că punctele nu sunt serios deviate de la linia dreaptă în cazul categoriei „Angajaţi permanent", ceea ce indică o distribuţie normală. În cazul grupei „Alte categorii de angajaţi", se constată deviaţii mari, evidenţiind abateri de la normalitate, fapt demonstrat şi cu ajutorul histogramei (vezi figura 6.16).

6.4.3 Procedeul P-P plot

Procedeul *P-P plot (Percent Percent)* compară funcţia de repartiţie a distribuţiei unei variabile empirice cu funcţia de repartiţie a unei distribuţii teoretice specificate (în cazul nostru, funcţia distribuţiei normale standard).

Construirea diagramei P-P plot presupune acelaşi demers prezentat pentru Q-Q plot, cu deosebirea că se alege din meniul *Graphs* comanda *P-P*.

Diagramele *P-P*, pentru exemplul dat, sunt prezentate în figura 6.19 şi evidenţiază aceleaşi situaţii ca diagramele Q-Q plot, prezentate în figura 6.18.

Observaţie! Procedeele grafice, aşa cum s-a constat din diagramele prezentate (histograma, Q-Q plot, P-P plot), *vizualizează diferenţele dintre o distribuţie empirică şi o distribuţie teoretică specificată.* Interpretarea lor se bazează pe intuiţie, fiind încărcate cu subiectivism.

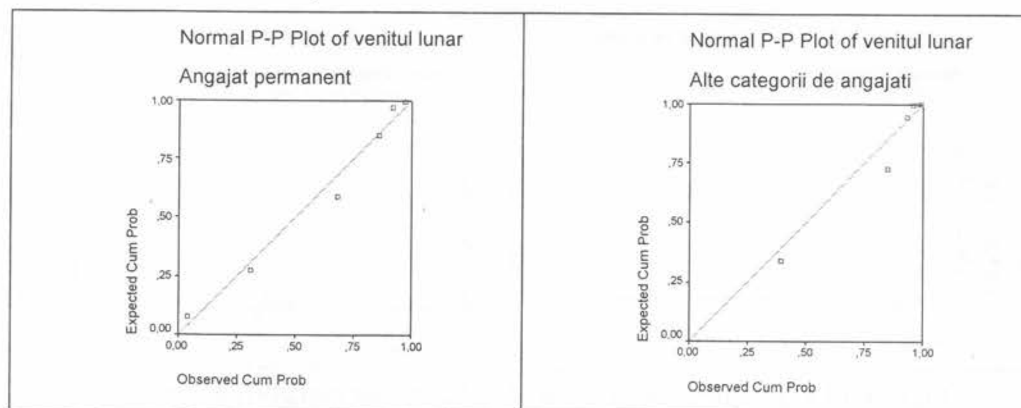


Figura 6.19 *P-P plot pentru variabila „Venit", pe categorii de angajați*

6.4.4 Procedee numerice (asimetria şi boltirea)

Ca procedee numerice pentru testarea normalităţii unei distribuţii, în SPSS sunt folosite *asimetria (Skewness)* şi *boltirea (Kurtosis)*, precum şi *testul Jarque-Bera*, *testul Shapiro-Wilk* şi *testul Kolmogorov-Smirnov*.

Asimetria şi boltirea

Asimetria (momentele centrate de ordin trei) şi boltirea (momentele centrate de ordin patru) arată în ce măsură distribuţia unei variabile deviază de la forma simetrică. Relaţiile de calcul sunt prezentate în paragraful 5.1.4.

Dacă o variabilă este distribuită normal, atunci are *asimetria egală cu zero* şi *boltirea egală cu trei*. Dacă asimetria este mai mare ca zero, distribuţia este asimetrică la dreapta, având mai multe observaţii în partea stângă a histogramei (vezi figura 6.16, diagrama „Alte categorii de angajaţi") şi invers în cazul unei asimetrii la stânga.

În SPSS, valorile asimetriei şi boltirii se obţin prin demersul: meniul *Analyze* → comanda *Descriptive Statistics* → opţiunea *Frequencies* (vezi paragraful 5.2.2). Rezultatele obţinute în output pentru exemplul dat (venitul pe categorii de angajaţi) sunt prezentate în figura 6.20.

Statistics[a]

venitul lunar al persoanei

N	Valid	160
	Missing	0
Mean		4,4500
Variance		5,9849
Skewness		1,251
Std. Error of Skewness		,192
Kurtosis		1,239
Std. Error of Kurtosis		,381

a. STPR2CAT = 1,00

Statistics[a]

venitul lunar al persoanei

N	Valid	240
	Missing	0
Mean		1,8167
Variance		3,9160
Skewness		3,053
Std. Error of Skewness		,157
Kurtosis		9,551
Std. Error of Kurtosis		,313

a. STPR2CAT = 2,00

Figura 6.20 *Output-ul obținut prin demersul:*
meniul Analyze → *comanda* Descriptive Statistics → *opțiunea* Frequencies

Valorile pentru asimetrie (*Skewness*) și boltire (*Kurtosis*) obținute pentru distribuția după variabila „venit" sunt diferite pentru cele două categorii de angajați, așa cum a reieșit și din procedeele grafice. Distribuția observată pentru angajații permanenți prezintă valori mici atât pentru asimetrie, cât și pentru boltire, relevând o distribuție normală, pe când în cazul altor categorii de angajați, valorile acestor statistici arată o asimetrie la dreapta pronunțată și o boltire cu abateri mari de la limitele normalității.

Observație! Deși statisticile asimetrie (*Skewness*) și boltire (*Kurtosis*) exprimă numeric în ce măsură o distribuție se abate de la normalitate, totuși nu dau posibilitatea interpretării gradului de semnificație a deviației de la normalitate.

6.4.5 Teste de normalitate (Jarque-Bera, Kolmogorov-Smirnov-Lilliefors)

Testul Jarque-Bera

Testul Jarque-Bera (*JB*) este fundamentat pe statistica ce urmează o lege χ^2 cu două grade de libertate, $JB \sim \chi^2_{v=2}$. Acest test cere să se verifice dacă valorile calculate ale coeficientului de asimetrie și ale coeficientului de boltire se abat de la valoarea 0, respectiv 3. Sub ipoteza de nul a normalității, valoarea așteptată a statisticii test este doi.

Testul de normalitate Jarque-Bera este definit de relația:

$$JB = n \cdot \left[\frac{\gamma_1^2}{6} + \frac{(\beta_2 - 3)^2}{24} \right],$$

unde:

γ_1 este coeficientul de asimetrie;
β_2 este coeficientul de boltire;
n este numărul de observații;
$6/n$ și $24/n$ reprezintă varianța asimetriei, respectiv a boltirii.

Se știe că momentele centrate de ordin impar ale unei distribuții normale sunt egale cu zero, iar momentul centrat de ordin patru este egal cu de trei ori σ^4.

Momentele centrate de ordin trei și de ordin patru sunt date de relațiile:

$$M(x_i - \mu)^3 = 0$$

$$M(x_i - \mu)^4 = 3\sigma^4.$$

Ca urmare, pentru o distribuție normală, coeficientul de asimetrie este egal cu zero, $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = 0$, iar coeficientul de boltire este egal cu trei, $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$.

Așadar, o distribuție normală este simetrică și mezocurtică.

Regula de decizie: dacă probabilitatea corespunzătoare valorii calculate a statisticii *JB* este superioară lui $\alpha = 0,05$, atunci se acceptă ipoteza de normalitate, H_0.

Acest test de normalitate este folosit în programul E-Views. În programul SPSS, acest test se calculează manual.

Testul Kolmogorov - Smirnov - Lilliefors (K-S-L)

Principiul verificării normalității unei distribuții pe baza acestui test constă în *compararea frecvențelor reale cumulate cu frecvențele teoretice cumulate extrase din tabelul Gauss.*

Ipoteza nulă presupune că cea mai mare diferență absolută dintre frecvențele cumulate ale valorii x_i a variabilei X observate nu depășește o anumită valoare extrasă din tabelul *K-S-L*, pentru un volum (n) dat și un risc admis.

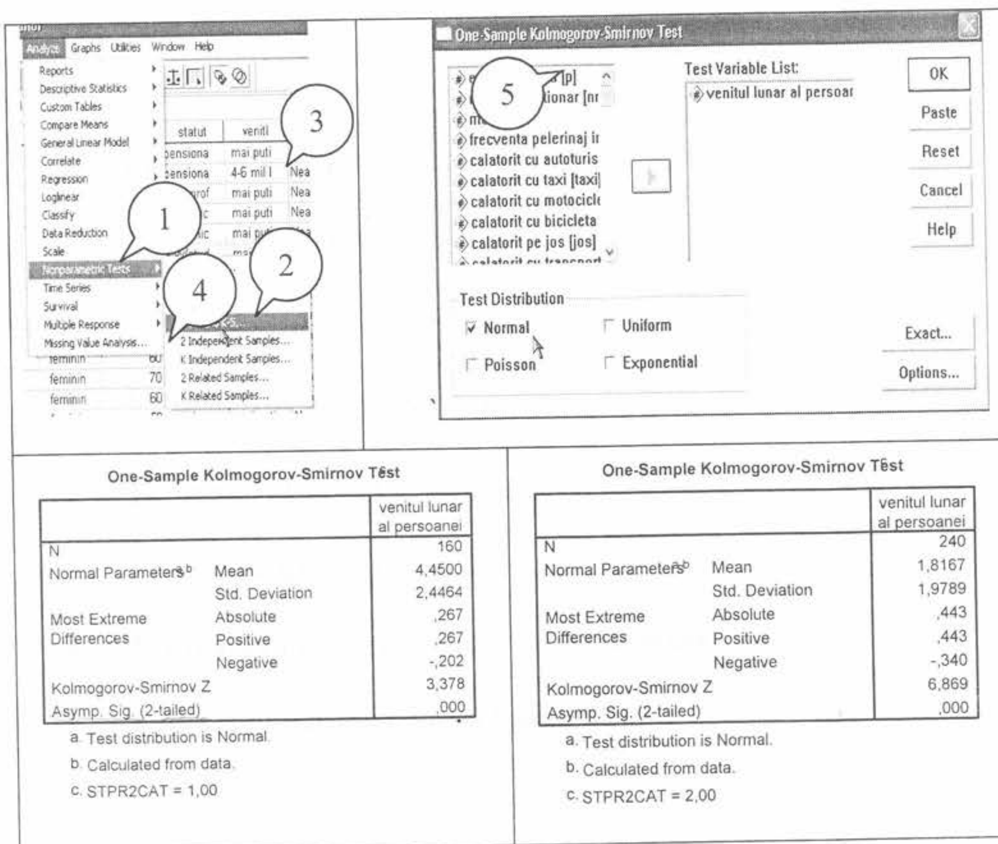
Verificarea normalității prin testul *K-S-L* presupune parcurgerea următorilor pași:

1. *Se calculează efectivele cumulate* (N_i);
2. *Se calculează valorile* p_i, adică ponderea efectivelor cumulate în totalul populației;
3. *Se află valorile* z_i corespunzătoare fiecărei valori x_i, pe baza valorilor $\overline{x}$ și s ale distribuției observate;

4. *Se citesc*, din tabelul Gauss-Laplace, *valorile teoretice* $p_{i'}$ corespunzătoare fiecărei valori z_i;
5. *Se calculează diferențele absolute* dintre valorile p_i și $p_{i'}$, adică dintre frecvențele reale cumulate p_i și frecvențele teoretice cumulate $p_{i'}$, și se alege cea mai mare diferență (în valoare absolută).

Admitem ipoteza de normalitate, adică ipoteza H_0 (ipoteza nulă), dacă la diferența maximă calculată găsim în tabelul *K-S-L* o valoare critică mai mare decât aceasta, pentru un volum dat al colectivității și un risc admis.

În SPSS, verificarea normalității cu ajutorul testului *K-S-L* presupune următorul demers: meniul *Analyze* → comanda *Nonparametric Test* → opțiunea *One-Sample Kolmogorov-Smirnov Test* (vezi figura 6.21).



One-Sample Kolmogorov-Smirnov Test

		venitul lunar al persoanei
N		160
Normal Parameters[a,b]	Mean	4,4500
	Std. Deviation	2,4464
Most Extreme Differences	Absolute	,267
	Positive	,267
	Negative	-,202
Kolmogorov-Smirnov Z		3,378
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.
b. Calculated from data.
c. STPR2CAT = 1,00

One-Sample Kolmogorov-Smirnov Test

		venitul lunar al persoanei
N		240
Normal Parameters[a,b]	Mean	1,8167
	Std. Deviation	1,9789
Most Extreme Differences	Absolute	,443
	Positive	,443
	Negative	-,340
Kolmogorov-Smirnov Z		6,869
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.
b. Calculated from data.
c. STPR2CAT = 2,00

Figura 6.21 Testul K-S-L *corespunzător aplicat la cele două distribuții ale variabilei „venit", pe categorii de angajați, și output-ul*

H_0: distribuția este normală $F(x) = \Phi(x)$
H_1: distribuția nu este normală $F(x) \neq \Phi(x)$
Sig ≥ α : Ac H_0
Sig < α : Resp. H_0 (Ac H_1)
Sig = 0,000, α = 0,05 } Sig < α ⇒ Ac H_1

Un nivel redus al gradului de semnificație (*Sig.* mai mic decât 0,05), așa cum a rezultat pentru exemplu dat (vezi figura 6.21), arată că distribuția diferă semnificativ de forma distribuției normale.

Aplicarea testului K-S-L în SPSS este posibilă și pe calea: meniul *Analyze* → comanda *Descriptive Statistics* → opțiunea *Explore*.

- În fereastra *Explore* se activează butonul de comandă *Plot*, care deschide fereastra *Explore: Plots*, unde se bifează caseta de validare a opțiunii *Normality plots with tests*;
- Prin butonul de comandă *Continue* se revine în fereastra dialog principală, unde, prin comanda *OK*, se cere ca SPSS să producă atât testul K-S-L, cât și normal Q-Q plot.

CAPITOLUL 7

ESTIMAREA PARAMETRILOR UNEI POPULAȚII

- **Probleme generale**
- **Proprietăți ale estimatorilor**
- **Estimatorul $\hat{\mu}$ al mediei μ**
- **Estimatorul $\hat{p}$ al proporției p**
- **Estimatorul $\hat{\sigma}^2$ al varianței σ^2**
- **Estimarea prin interval de încredere**
- **Estimarea mediei prin interval de încredere**
- **Estimarea parametrilor folosind SPSS**

Datele folosite în procesul cunoaşterii statistice sunt, de regulă, rezultatul observării unui eşantion reprezentativ extras dintr-o populaţie, foarte rar populaţia ţintă putând fi înregistrată în întregime, gen recensământ.

Obiectivul cunoaşterii fiind populaţia, rezultatele observate pe un eşantion sunt generalizate la nivelul populaţiei prin *estimare statistică*. Incertitudinea inerentă unei astfel de cunoaşteri se exprimă utilizând teoria probabilităţilor.

Primele contribuţii la teoria estimaţiei au fost aduse de A.M. Legendre (1805), C.F. Gauss (1809), P.S. Laplace, K. Pearson, R.A. Fisher (*Teoria statistică a estimaţiei*, 1925).

7.1 Probleme generale

Problemele generale ale unei estimări statistice vizează:
- precizarea noţiunilor folosite;
- definirea distribuţiei de selecţie a estimatorilor;
- definirea teoremelor ce stau la baza estimării.

7.1.1 Noţiuni şi termeni pereche

Estimare. Prin estimare se înţelege un procedeu prin care se generalizează rezultatele observate pe un eşantion, la nivelul populaţiei din care este extras, adică se află valoarea unui parametru al unei populaţii pe baza datelor înregistrate la nivelul unui eşantion extras din aceasta.

Estimarea se poate efectua fie sub formă de *estimare punctuală*, fie sub formă de *estimare prin interval de încredere.*

Estimarea punctuală presupune estimarea unei *valori posibile* a estimatorului parametrului căutat, adică o estimaţie calculată pe baza datelor înregistrate la nivelul unui eşantion.

Estimarea prin interval de încredere presupune aflarea limitelor de încredere ale unui interval care acoperă *valoarea adevărată a unui parametru* al populaţiei. Estimarea prin interval de încredere ţine seama de fluctuaţiile distribuţiei de selecţie a estimatorului parametrului considerat.

În estimarea statistică se folosesc o serie de termeni pereche: *parametri, estimatori, estimaţii* (vezi tabelul 7.1).

Parametri. În procesul estimării, un parametru reprezintă o *mărime fixă, reală*, dar necunoscută a unei populații. *Parametrul este „valoarea reală" care trebuie estimată.* Parametrul se notează printr-o literă din alfabetul grec, în general prin θ, și se determină pe baza unei funcții (medie, varianță etc.) a caracteristicii X observate la nivelul unei populații.

Estimatori (statistici). Un estimator este o *statistică* (o variabilă aleatorie) ce urmează o lege de probabilitate care depinde, în general, de un parametru necunoscut și este utilizat pentru a estima un parametru al populației.

Estimatorul se notează cu aceeași literă folosită pentru un parametru, adăugând o pălărie (accent circumflex) deasupra, sau cu litere din alfabetul latin.

Pentru un parametru θ, un estimator al său se notează cu $\hat{\theta}$ și reprezintă o funcție de n variabile aleatorie de selecție$(X_1, X_2, ..., X_n)$ independente și identic distribuite:

$$\hat{\theta} = f(X_1, X_2, ..., X_n).$$

Estimații (valori tipice de sondaj). O estimație reprezintă o valoare $\hat{\theta}_i$ a unui estimator $\hat{\theta}$ al parametrului θ. Este calculată pe baza unui eșantion de n valori $(x_1, x_2, ..., x_n)$, prelevat dintr-o populație N, adică este o *valoare tipică* de sondaj de forma:

$$f(x_1, x_2, ..., x_n).$$

Tabelul 7.1 *Termeni folosiți în procesul de estimare statistică*

	Parametri (valori)	**Estimatori (variabile)**	**Estimații (valori)**
1. Media	μ	$\hat{\mu}$	$\bar{x}$
2. Varianța	σ^2	$\hat{\sigma}^2$	s^2, s'^2
3. Proporția	p	$\hat{p}$	f
4. Coeficienții de regresie	α, β	$\hat{\alpha}$, $\hat{\beta}$	a, b
5. Coeficientul de corelație	ρ	$\hat{\rho}$	$r_{y,x}$

Observație! Termenii pereche (parametri – estimații sau valori tipice de sondaj) întâlniți în procesul de estimare au același conținut metodologic, dar se deosebesc din punctul de vedere al informațiilor folosite. Astfel, media și varianța, valori tipice (valori caracteristice) pentru cele două categorii de colectivități (populația și eșantionul), sunt numite diferit – *parametri* pentru

populație și *valori tipice de sondaj* pentru eșantion – datorită caracterului lor diferit pe care îl au într-o cercetare prin sondaj. Atât parametrii, cât și valorile tipice de sondaj sunt valori reale, calculate pe baza datelor observate la nivelul unei populații, respectiv la nivelul unui eșantion. De regulă, parametrii unei populații nu pot fi calculați direct, nefiind posibilă observarea întregii populații. Parametrii sunt estimați pe baza valorilor tipice de sondaj, rezultate din prelucrarea datelor unui eșantion.

7.1.2 Distribuții de selecție

O distribuție de selecție este distribuția unei statistici, $\hat{\theta}$. Dacă variabila aleatorie este media de selecție, atunci ne aflăm în cazul distribuției mediei de selecție, iar dacă variabila este proporția de selecție, respectiv varianța de selecție, este cazul distribuției proporției de selecție, respectiv al varianței de selecție.

Dintr-o populație de volum N, de parametri μ (media populației) și σ^2 (varianța sa), se pot extrage k eșantioane de volum n.

Astfel:

- în cazul eșantionării cu revenire (repetată):

$$k = N^n;$$

- în cazul eșantionării aleatorie fără revenire (nerepetată), k este dat de numărul combinărilor de N elemente ale populației luate câte n:

$$k = C_N^n = \frac{N!}{(N-n)!\,n!}$$

Fiecare eșantion are media ($\bar{x}_j$), proporția (f_j) și varianța sa (s_j^2). Pe ansamblul celor k eșantioane se obțin variabilele:

$$\hat{\mu} : (\bar{x}_1, \bar{x}_2, ..., \bar{x}_k);$$

$$\hat{p} : (f_1, f_2, ..., f_k);$$

$$\hat{\sigma}^2 : (s_1^2, s_2^2, ..., s_k^2),$$

unde:

$\hat{\mu}$ – *media de selecție*;

$\hat{p}$ – *proporția de selecție*;

$\hat{\sigma}^2$ – *varianța de selecție*.

Valorile posibile ale fiecărei variabile se abat mai mult sau mai puțin de la valoarea parametrului corespunzător colectivității generale (vezi figura 7.1).

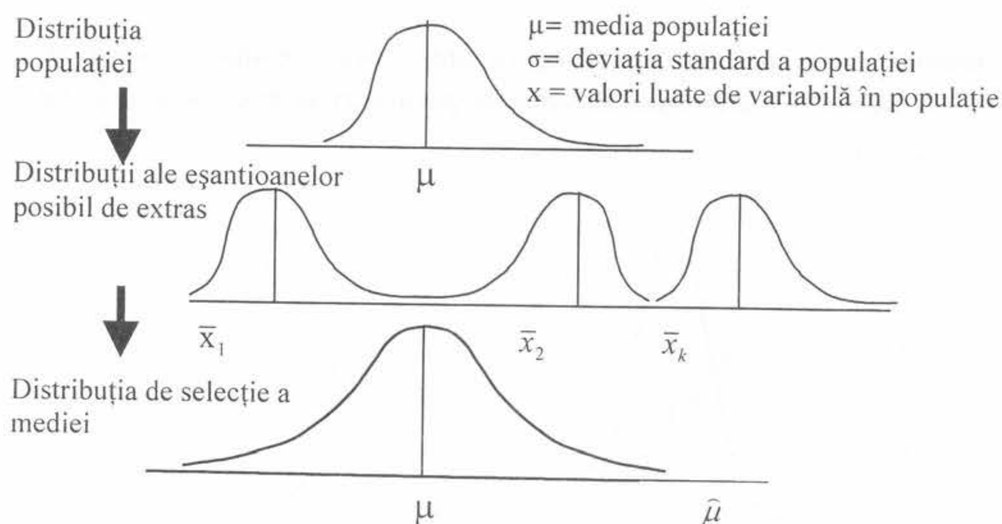


Figura 7.1 *Distribuția unei populații, distribuții pentru k eșantioane, distribuția de selecție a mediei*

Media $\hat{\mu}$, numită *medie de selecție*, proporția $\hat{p}$, numită *proporție de selecție*, și varianța $\hat{\sigma}^2$, numită *varianță de selecție*, apar ca variabile, cu niveluri diferite pentru fiecare eșantion (vezi și figura 7.1).

Fiecărui nivel al mediei de selecție, al proporției de selecție, al varianței de selecție îi corespunde o anumită frecvență de apariție. Frecvența de apariție a mediei unui eșantion, de exemplu, poate fi interpretată ca probabilitate de apariție a acesteia. În cazul extragerii tuturor eșantioanelor posibile, suma tuturor probabilităților de apariție a mediei este egală cu *1*.

Pentru fiecare variabilă, se determină un nivel mediu $M(\hat{\theta})$ și o varianță $V(\hat{\theta})$, notată și $\sigma^2_{\hat{\theta}}$.

7.2 Proprietăți ale estimatorilor

Principalele proprietăți ale estimatorilor sunt: *nedeplasare*, *convergență* și *eficiență*.

7.2.1 Nedeplasare

Un estimator este nedeplasat, adică fără bias, dacă speranța matematică a variabilei aleatorii $\hat{\theta}$ *este egală cu valoarea parametrului* θ *din populație* (vezi figura 7.2): $M(\hat{\theta}) = \theta$.

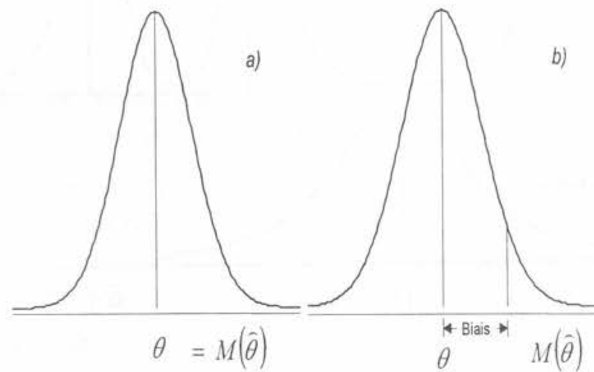


Figura 7.2 *Estimator nedeplasat (a) și estimator deplasat (b)*

7.2.2 Convergență

Un estimator este convergent (consistent) dacă varianța sa $V(\hat{\theta})$ *tinde spre zero* ($V(\hat{\theta}) \to 0$), când volumul eșantionului tinde spre volumul populației, adică:

$$\lim_{n \to N} P(|\hat{\theta} - \theta| < \varepsilon) = 1.$$

Varianța $V(\hat{\theta})$ măsoară incertitudinea care planează asupra *calității estimatorului.*

Un estimator este *corect* dacă îndeplinește condițiile:

$$\left.\begin{matrix} M(\hat{\theta}) \to \theta \\ V(\hat{\theta}) \to 0 \end{matrix}\right\} \text{ atunci când } n \to N.$$

Un estimator este *absolut corect* dacă îndeplinește următoarele două condiții:

$M(\hat{\theta}) = \theta$ (estimator nedeplasat);

$V(\hat{\theta}) \to 0$, dacă $n \to N$ (estimator convergent).

7.2.3 Eficiență

Un estimator este considerat eficient dacă este convergent și are *varianța cea mai mică posibil* față de varianța oricărui alt estimator calculat pentru același eșantion de volum n (vezi figura 7.3 b): $V(\hat{\theta}) = minim$.

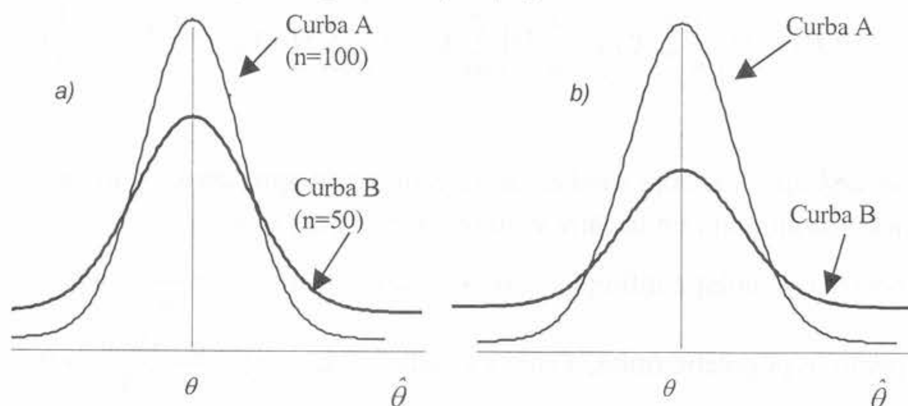


Figura 7.3 *Estimator convergent (a) și estimator eficient (b)*

Observație! În figura 7.3a se observă că varianța estimatorului tinde să se reducă o dată cu mărirea volumului eșantionului. Analog, figura 7.3b prezintă distribuția comparativă a doi estimatori cu grade diferite de eficiență (curba A arată distribuția mediei de selecție, iar curba B arată distribuția medianei de selecție).

7.3 Estimatorul $\hat{\mu}$ al mediei μ

Media μ a populației se poate estima punctual prin media ($\bar{x}$) obținută la nivelul unui eșantion. Media $\bar{x}$ este o valoare a estimatorului $\hat{\mu}$, calculată pe baza datelor de la nivelul unui eșantion.

7.3.1 Proprietățile estimatorului $\hat{\mu}$

Estimatorul mediei urmează legea normală. Este un estimator nedeplasat, convergent și eficient.

Nedeplasare. Dacă parametrul este μ, media populației, și $\hat{\mu}$ este estimatorul său, atunci se poate arăta că $\hat{\mu}$ este un *estimator nedeplasat*:

$$M(\hat{\mu}) = \mu;$$

$$M(\hat{\mu}) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} M\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n} n \cdot \mu = \mu$$

Convergență. Varianța mediei de selecție tinde spre zero, $V(\hat{\mu}) \to 0$, când volumul eşantionului tinde către volumul populației ($n \to N$):

– pentru o populație infinită, $V(\hat{\mu}) \to 0$, adică $V(\hat{\mu}) = \sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n} \to 0$;

– pentru o populație finită, $V(\hat{\mu}) \to 0$, adică $V(\hat{\mu}) = \sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N} \to 0$.

7.3.2 Distribuția mediei de selecție. Teorema limită centrală

Distribuția mediei de selecție se fundamentează pe teorema limită centrală *(TLC)*, conform căreia *suma unui număr suficient de mare de variabile (X) aleatorii independente şi identic distribuite urmează aproximativ o lege normală.*

Distribuția mediei de selecție tinde spre o distribuție normală:

- când volumul eşantionului, *n*, tinde spre infinit, *indiferent de legea de distribuție urmată de variabila aleatorii de distribuție a populației de origine* (vezi figura 7.4).;
- când volumul eşantionului este oricât de mic, dacă variabila aleatorie de distribuție a populației urmează o distribuție aproximativ normală.

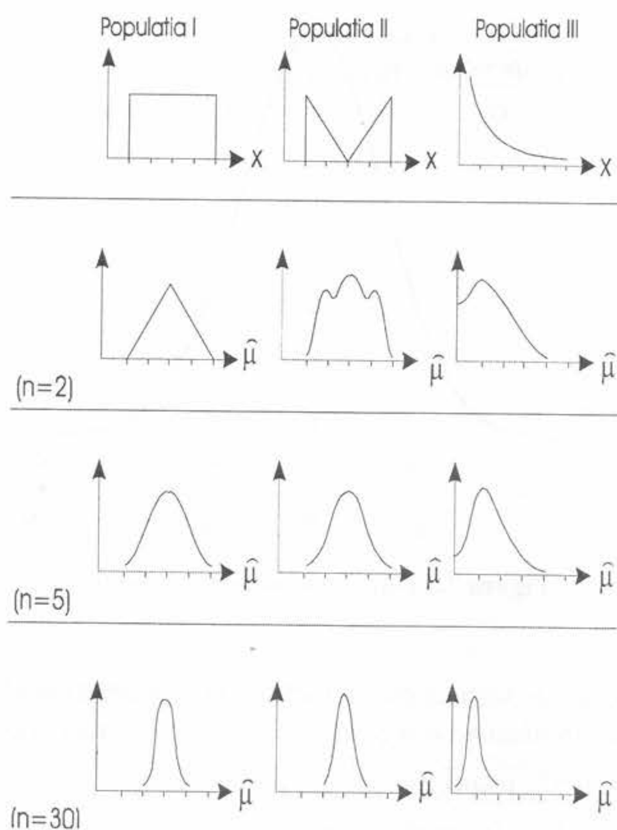


Figura 7.4 *Tendința spre normalitate a distribuției mediei de selecție în raport cu mărirea volumului eșantionului (TLC)*

Forma distribuției mediei de selecție ia alura unei curbe normale, abaterile într-un sens sau altul față de media lor compensându-se reciproc.

În figura 7.5, $\sigma_{\hat{\mu}}$ reprezintă abaterea standard a estimatorului $\hat{\mu}$ al parametrului μ.

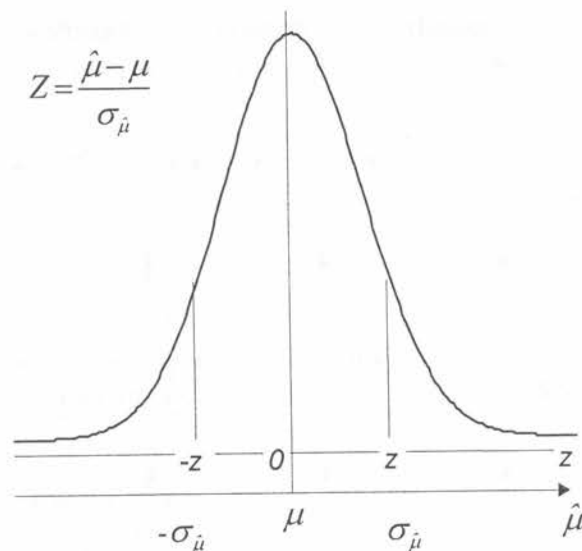


Figura 7.5 *Distribuţia mediei de selecţie*

Distribuţia mediei de selecţie este caracterizată prin următoarele:

1. $\hat{\mu}$ *urmează întotdeauna o lege normală* sau aproximativ normală, de medie μ şi varianţă $\sigma^2_{\hat{\mu}}$, respectiv: $\hat{\mu} \sim N(\mu, \sigma^2_{\hat{\mu}})$;
2. *media distribuţiei mediei de selecţie* este egală cu media populaţiei: $M(\hat{\mu}) = \mu$;
3. *varianţa mediei de selecţie* ($\sigma^2_{\hat{\mu}}$) este egală cu varianţa populaţiei (σ^2) împărţită la volumul eşantionului (n) şi se calculează după relaţiile:

- pentru o populaţie infinită (cazul sondajului aleatoriu repetat):

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n};$$

- pentru o populaţie finită (cazul sondajului aleatoriu nerepetat),

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n}\frac{N-n}{N},$$

unde:

σ^2 – varianţa populaţiei;
n – volumul eşantionului;
N – volumul populaţiei.

Varianța mediei de selecție stă la baza calculului *erorii medii de reprezentativitate*, care se măsoară în unități de abateri standard, adică prin abaterea medie pătratică a mediei de selecție față de media ei ($\sigma_{\hat{\mu}}$).

Estimatorul varianței mediei de selecție ($\hat{\sigma}^2_{\hat{\mu}}$) este definit de relațiile:

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{\hat{\sigma}'^2}{n}, \text{ respectiv } \hat{\sigma}^2_{\hat{\mu}} = \frac{\hat{\sigma}'^2}{n}\frac{N-n}{N},$$

unde $\hat{\sigma}'^2$ reprezintă un estimator al *varianței de selecție corectată (estimator nedeplasat al varianței populației)*.

O estimație nedeplasată a varianței mediei de selecție se determină prin relația:

$$s'^2_{\hat{\mu}} = \frac{s'^2}{n}, \quad unde \quad s'^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Observație! Din relațiile prezentate rezultă că *mărimea varianței mediei de selecție este direct proporțională cu varianța colectivității generale și invers proporțională cu volumul eșantionului.*

Dacă dorim să mărim sau să micșorăm varianța mediei de selecție, va trebui să micșorăm sau să mărim volumul eșantionului (n) cu o constantă K, astfel:

– pentru mărire: $K\sigma^2_{\hat{\mu}} = \dfrac{\sigma^2}{\frac{1}{K}n}$; respectiv $K\sigma_{\hat{\mu}} = \sqrt{\dfrac{\sigma^2}{\frac{1}{K^2}n}}$;

– pentru micșorare: $\dfrac{\sigma^2_{\hat{\mu}}}{K} = \dfrac{\sigma^2}{Kn}$, respectiv $\dfrac{\sigma_{\hat{\mu}}}{K} = \sqrt{\dfrac{\sigma^2}{K^2 n}}$.

7.4 Estimatorul $\hat{p}$ al proporției p

7.4.1 Definiție

Estimatorul proporției p se notează cu $\hat{p}$. Este o variabilă aleatorie care urmează o lege Bernoulli de medie $p = \dfrac{N_A}{N}$, unde N_A reprezintă indivizii din categoria

A, iar N – întreaga populație. O valoare posibilă a estimatorului $\hat{p}$ este proporția (f) calculată la nivelul unui eșantion. Reprezintă o *estimație* definită prin:

$$f = \frac{n_A}{n},$$

unde:

n_A este numărul unităților din categoria A din eșantion;
n este volumul eșantionului.

7.4.2 Proprietăți ale estimatorului $\hat{p}$

Estimatorul proporției $\hat{p}$ este un *estimator nedeplasat* și *convergent* al proporției p.

Se demonstrează că:

1. $M(\hat{p}) = p$, indiferent dacă selecția este fără revenire (nerepetată) sau nu;

2. $V(\hat{p}) = \sigma^2_{\hat{p}} = \dfrac{p(1-p)}{n}$, dacă populația este infinită,

$$V(\hat{p}) = \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1},$$ dacă populația este finită și $n \rightarrow N$.

Varianța proporției de selecție ($\sigma^2_{\hat{p}}$) stă la baza calculului *erorii medii de reprezentativitate a proporției* ($\sigma_{\hat{p}}$).

Estimatorul varianței proporției de selecție $\hat{\sigma}^2_{\hat{p}}$ este dat de relațiile:

$\hat{\sigma}^2_{\hat{p}} = \dfrac{\hat{p}(1-\hat{p})}{n-1}$, în cazul populației infinite sau a selecției repetate, respectiv,

$\hat{\sigma}^2_{\hat{p}} = \dfrac{\hat{p}(1-\hat{p})}{n-1} \cdot \dfrac{N-n}{N}$, în cazul selecției nerepetate.

7.5 Estimatorul $\hat{\sigma}^2$ al varianței σ^2

7.5.1 Estimarea punctuală a varianței σ^2

Estimarea punctuală a varianței σ^2 presupune calculul unei estimații s^2 pe baza datelor unui eșantion prelevat din populație.

Dacă parametrul este varianța populației, σ^2, atunci $\hat{\sigma}^2$ este estimatorul său și este definit prin relația: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.

O *estimație* calculată la nivelul unui eșantion de volum n se poate afla după relația: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Estimatorul $\hat{\sigma}^2$ este un *estimator deplasat*. Media estimatorului $\hat{\sigma}^2$ este definită de relația:

$$M(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2.$$

Se observă că media estimatorului $\hat{\sigma}^2$ este diferită de σ^2, ceea ce semnifică faptul că $\hat{\sigma}^2$ este un estimator deplasat.

Bias-ul estimatorului este: $B(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$.

7.5.2 Estimatorul varianței distribuției de selecție a diferenței dintre două medii și a diferenței dintre două proporții

Estimatorul dintre 2 medii

a) *Când varianțele a două populații comparate sunt diferite între ele*, $\sigma_1^2 \neq \sigma_2^2$, varianța de selecție a diferenței dintre două medii se calculează după relația:

$$\sigma^2_{\hat{\mu}_1 - \hat{\mu}_2} = \sigma^2_{\hat{\mu}_1} + \sigma^2_{\hat{\mu}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Estimatorul său este:

$$\hat{\sigma}'^2_{\hat{\mu}_1-\hat{\mu}_2} = \frac{\hat{\sigma}'^2_1}{n_1} + \frac{\hat{\sigma}'^2_2}{n_2}.$$

b) *Când varianțele celor două populații comparate sunt egale între ele* ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), estimatorul varianței de selecție a diferenței dintre două medii este:

$$\hat{\sigma}'^2_{\hat{\mu}_1-\hat{\mu}_2} = \hat{\sigma}^2_w\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

unde:

$$\hat{\sigma}^2_w = \frac{\hat{\sigma}'^2_1(n_1 - 1) + \hat{\sigma}'^2_2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Estimația varianței de selecție a diferenței dintre două medii se calculează după relația:

$$s'^2_{\hat{\mu}_1-\hat{\mu}_2} = s^2_w\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

unde s_w^2 reprezintă varianța ponderată obținută pe baza estimațiilor varianțelor calculate la nivelul eșantioanelor n_1, n_2 și se află după relația:

$$s_w^2 = \frac{s'^2_1(n_1 - 1) + s'^2_2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Estimatorul varianței de selecție a diferenței dintre două proporții

Varianța de selecție a diferenței dintre două proporții se calculează în aceeași manieră ca varianța diferenței dintre două medii, după relația:

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2},$$

unde p_1 și p_2 sunt proporțiile populațiilor comparate.

Estimatorul corespunzător este:

$$\hat{\sigma}^2_{\hat{p}_1-\hat{p}_2} = \hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2} = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}.$$

7.6 Estimarea prin interval de încredere

7.6.1 Situații

Estimarea parametrului θ se face pe baza estimatorului $\hat{\theta}$. Deoarece estimatorul este o variabilă aleatorie, este necesară cunoaşterea legii de distribuție a acestuia.

În funcție de cunoaşterea sau necunoaşterea legii de distribuție a estimatorului, în procesul de estimare prin interval de încredere se întâlnesc trei situații, definite în funcție de volumul eşantionului şi de cunoaşterea sau nu a legii de distribuție a populației.

1. *Legea populației este cunoscută.* În acest caz, în studiul comportamentului estimatorului se consideră legea populației.
2. *Legea populației este cunoscută şi eşantionul este de volum mare.* Când eşantionul este de volum mare, se utilizează comportamentul asimptotic, legea estimatorului tinde spre legea normală.
3. *Legea populației nu este cunoscută.* Comportamentul estimatorului este *a priori* necunoscut. În acest caz, se recurge la teoreme limită pentru a obține o lege asimptotică.

Estimarea prin interval de încredere constă în căutarea unui interval în care probabil se situează valoarea unui parametru necunoscut din populația totală. Valoarea estimată a parametrului este influențată de fluctuațiile de selecție, valoarea sa depinzând de valorile statistice ale eşantionului extras. În estimarea prin interval de încredere, se pleacă de la o *estimație punctuală* obținută prin observarea unui eşantion şi de la definirea *limitelor de încredere ale intervalului* care acoperă cu o anumită probabilitate valoarea adevărată, dar necunoscută a unui parametru, pentru un coeficient de încredere dat.

7.6.2 Intervalul de încredere (I.C.)

A defini un interval de încredere înseamnă *a căuta limitele de încredere,* $L_i = \hat{\theta} - \Delta_{\hat{\theta}}$ *şi* $L_s = \hat{\theta} + \Delta_{\hat{\theta}}$, *care acoperă valoarea parametrului* θ, *pentru un coeficient de încredere*: $P(L_i \leq \theta \leq L_s) = 1 - \alpha$, adică:

$$\text{I.C.} = [\hat{\theta} - \Delta_{\hat{\theta}}; \hat{\theta} + \Delta_{\hat{\theta}}],$$

unde:

L_i *şi* L_s – limitele de încredere: inferioară, respectiv superioară;

($1-\alpha$) – probabilitatea cu care se garantează că intervalul acoperă valoarea adevărată a parametrului θ, în cazul unei probleme de estimație, respectiv valoarea unei statistici, în cazul unei probleme de distribuție de selecție;

α – riscul, respectiv probabilitatea ca intervalul să nu conțină valoarea căutată.

Calculul limitelor intervalului de încredere pornește de la estimarea valorii erorii limită, $\Delta_{\hat{\theta}}$, pe baza distribuției de selecție a estimatorului $\hat{\theta}$ al parametrului θ. Determinarea erorii limită este necesară pentru realizarea unei estimări prin interval de încredere.

7.6.3 Eroarea limită

Eroarea limită se determină pentru un estimator ($\hat{\theta}$) al unui parametru (θ), ținându-se seama, pe de o parte, de legea de distribuție a acestuia, iar pe de altă parte, de mărimea erorii medii de selecție corespunzătoare tipului de sondaj practicat. Eroarea limită se calculează ca *produs între coeficientul de încredere al unei legi de distribuție a unui estimator și eroarea medie de reprezentativitate a acestuia.*

Legea de distribuție specifică distribuției de selecție a mediei este, conform *TLC*, o lege normală.

Eroarea limită a mediei de selecție se calculează după relațiile:

$\Delta_{\hat{\mu}} = z \cdot \sigma_{\hat{\mu}}$, în cazul în care se cunoaște parametrul σ, respectiv:

$\Delta_{\hat{\mu}} = t \cdot \sigma_{\hat{\mu}}$, în cazul în care se estimează parametrul σ.

Mărimea erorii medii $\sigma_{\hat{\mu}}$ se calculează diferențiat, așa cum s-a prezentat în paragrafele anterioare, în funcție de estimatorul considerat și de tipul de sondaj practicat.

7.7 Estimarea mediei prin interval de încredere

Dacă parametrul căutat este μ – media unei populații, iar $\hat{\mu}$ – media de selecție, construirea I.C. pleacă de la o ipoteză asupra distribuției mediilor de selecție, deci și a abaterilor medii pătratice a acestora față de media populației, respectiv față de media lor.

Estimarea mediei prin I.C., în cazul în care se cunoaște legea populației și aceasta este o lege normală, $N(\mu, \sigma^2)$, poate prezenta două situații:

- *când se cunoaște varianța;*
- *când nu se cunoaște varianța.*

7.7.1 Construirea intervalului de încredere când se cunoaște varianța unei populații

În acest caz, *legea estimatorului* $\hat{\mu}$ este o lege normală:

$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{n})$, respectiv, sub formă redusă, $\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Construirea intervalului de încredere se bazează pe variabila normală centrată redusă Z:

$$Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}.$$

Această variabilă permite să se construiască un interval de încredere:

$$P(-z_{\alpha/2} \leq \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha,$$

unde:

$z_{\alpha/2}$ este o valoare a variabilei normale centrate reduse Z,

α este un nivel al probabilității, cuprins între zero și unu.

La nivelul unui eșantion ($\hat{\mu}$ ia valoarea $\bar{x}$), intervalul este:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

cu o încredere ($1-\alpha$). Valorile sunt simetrice.

Construirea intervalului de încredere al mediei, *când* σ *este cunoscută*, este prezentată în figura 7.6.

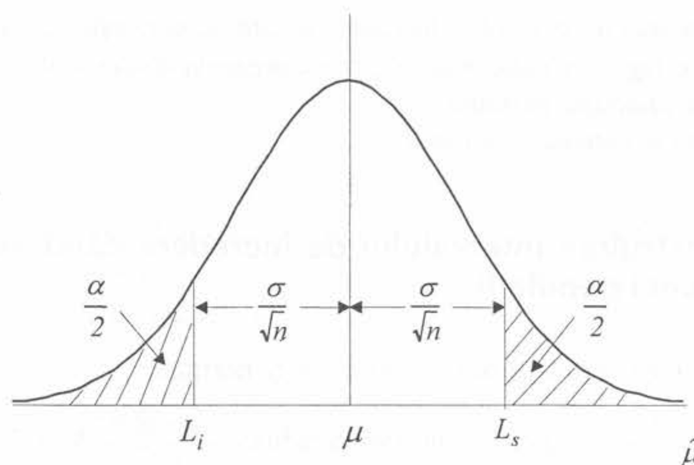


Figura 7.6 *Intervalul de încredere pentru medie, cu σ cunoscută*

Calculul şi interpretarea I.C. se face în funcţie de valorile luate de variabila *Z*.

Astfel,

– pentru (*Z* = *1*), I.C. este:

$(\bar{x} - 1 \cdot \sigma_{\hat{\mu}}) < \mu < (\bar{x} + 1 \cdot \sigma_{\hat{\mu}})$;

– pentru (*Z* = *2*), I.C. este:

$(\bar{x} - 2 \cdot \sigma_{\hat{\mu}}) < \mu < (\bar{x} + 2 \cdot \sigma_{\hat{\mu}})$.

Observaţie! În condiţiile unei *erori medii de reprezentativitate date* ($\sigma_{\hat{\mu}}$), cu un *Z* = *1*, putem spera ca valoarea estimată a mediei populaţiei să coincidă cu valoarea adevărată a acesteia, în medie, în *68* din *100* de ocazii; pentru *Z* = *2*, ar exista, în medie, *95* de şanse din *100* ca acest eveniment să se întâmple ş.a.m.d.

7.7.2 Construirea intervalului de încredere când nu se cunoaşte varianţa unei populaţii

Deoarece varianţa (σ) *nu este cunoscută*, ea se înlocuieşte cu o estimaţie a sa, s'^2, abaterea medie pătratică corectată a eşantionului observat.

S-a demonstrat că:

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}' / \sqrt{n}} \sim t(n-1),$$

unde:

$t(n-1)$ este legea de distribuție *Student* de $(n-1)$ grade de libertate.

Pentru $P(t > t_{\alpha/2}) = \alpha / 2$, la nivelul unui eşantion, vom avea:

$t = \frac{\bar{x} - \mu}{s' / \sqrt{n}}$, iar intervalul va fi:

$\left(\bar{x} \pm t_{\alpha/2} \frac{s'}{\sqrt{n}}\right)$, cu un grad de încredere $(1-\alpha)$.

7.8 Estimarea parametrilor folosind SPSS

Calculul *intervalului de încredere* pentru o medie sau pentru o proporție presupune efectuarea următoarelor operații:
- calculul valorii tipice de sondaj (media eşantionului, de exemplu);
- determinarea variabilității estimatorului considerat (varianța mediei de selecție, de exemplu);
- alegerea nivelului de încredere (90%, 95%, 99%);
- calculul limitelor intervalului de încredere.

7.8.1 Estimarea mediei

SPSS calculează *valoarea tipică de sondaj*, în cazul nostru media eşantionului ($\bar{x}$), *scorul Z* corespunzător şi *eroarea standard a mediei* ($\sigma_{\hat{\mu}}$), precum şi *limita inferioară* şi *limita superioară ale intervalului de încredere.*

Paşii de urmat sunt (vezi figura 7.7):
- selectăm succesiv: meniul *Analyze* → comanda *Descriptive Statistics* → opțiunea *Explore*;
- în fereastra *Explore*, selectăm variabila dorită (de exemplu, vârsta) şi o mutăm în zona *Dependent List*;
- activăm butonul de comandă *Statistics* care deschide fereastra *Explore: Statistics*, unde bifăm caseta de validare *Descriptives* şi precizăm în

caseta *Confidence Interval for Mean* nivelul de încredere dorit (de obicei, 90%, 95% sau 99%). Valoarea implicită este 95%;

- butonul de comandă *Continue* determină revenirea în fereastra *Explore*, din care activăm *OK* pentru a comanda afişarea rezultatelor în fereastra *Output* (vezi figura 7.8).

Explore
mod placut de a cala
mod placut de a cala
mod placut de a cala
protejeaza mediul [e
protejeaza mediul m
protejeaza mediul ai
sexul persoanei [se
statutul socio-profes
venitul lunar al pers
judetul in care locuie
rpvarsta
lgvarsta
Dependent List:
varsta persoanei [varst
Factor List:
Display
Both Statistics Plots
Statistics...
OK
Paste
Reset
Cancel
Explore: Statistics
Descriptives
Confidence Interval for Mean: 95 %
M-estimators
Outliers
Percentiles
Continue Cancel Help
1 2 3

Figura 7.7 *Ferestrele* Explore *şi* Explore: Statistics

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
varsta persoanei	400	100,0%	0	,0%	400	100,0%

Descriptives

			Statistic	Std. Error
varsta persoanei	Mean		38,03	,75
	95% Confidence Interval for Mean	Lower Bound	36,55	
		Upper Bound	39,52	
	5% Trimmed Mean		37,43	
	Median		40,00	
	Variance		226,976	
	Std. Deviation		15,07	
	Minimum		17	
	Maximum		70	
	Range		53	
	Interquartile Range		28,00	
	Skewness		,468	,122
	Kurtosis		-,822	,243

Media
Eroarea standard
Limita inferioară
Limita superioară

Figura 7.8 *Rezultate SPSS pentru variabila „Vârsta" din* tapestry.sav.

IC $36,55 \le \mu \le 39,52$
$\bar{x} = 38,03$ $s_\mu = 0,75$
$IC = \bar{x} \pm z_{\alpha/2} \cdot s_\mu = 38,03 \pm 1,96 \cdot 0,75$ (36,55; 39,52)
$\alpha = 0,05 \Rightarrow z_{\alpha/2} = 1,96$

Putem spune cu o încredere de 95% că vârsta medie a populației este între 36,55 și 39,52 ani.

Observație! Dacă se modifică nivelul de încredere, atunci se constată că se schimbă și limitele intervalului de încredere (vezi figura 7.9).

			Statistic	Std. Error
varsta persoanei	Mean		38,03	,75
	90% Confidence Interval for Mean	Lower Bound	36,79	
		Upper Bound	39,28	

			Statistic	Std. Error
varsta persoanei	Mean		38,03	,75
	95% Confidence Interval for Mean	Lower Bound	36,55	
		Upper Bound	39,52	

			Statistic	Std. Error
varsta persoanei	Mean		38,03	,75
	99% Confidence Interval for Mean	Lower Bound	36,09	
		Upper Bound	39,98	

Figura 7.9 *Intervale de încredere pentru variabila „Vârsta" la 90%, 95% și 99%, obținute prin demersul: meniul* Analyze → *comanda* Descriptive Statistics → *opțiunea* Explore

Aceleași rezultate se obțin urmând demersul: meniul *Analyze* → comanda *Compare Means* → opțiunea *One-Sample T Test*, după care se parcurg pașii:

- în fereastra de dialog *One-Sample T Test* (vezi figura 7.10), selectăm variabila *vârsta* și o mutăm în zona *Test Variable(s)*;
- acceptăm valoarea implicită *0* în caseta *Test Value*;
- activăm butonul de comandă *OK* și SPSS calculează intervalul de încredere pentru 95% (vezi figura 7.11).

→ Pt o încredere de 95%, vârsta medie a populației din care s-a extras eșantionul este cuprinsă între 36,55 și 39,52 ani

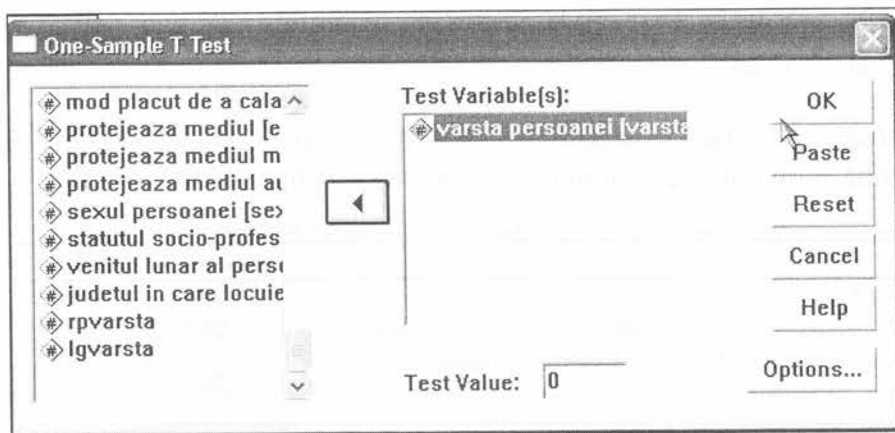


Figura 7.10 *Fereastra de dialog* One-Sample T Test

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
varsta persoanei	400	38,03	15,07	,75

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
varsta persoanei	50,492	399	,000	38,03	36,55	39,52

Figura 7.11 *Intervalul de încredere pentru media variabilei „Vârsta", calculat prin demersul: meniul* Analyze → *comanda* Compare Means → *opțiunea* One-Sample T Test

7.8.2 Estimarea proporției

În SPSS nu este calculat direct intervalul de încredere pentru o proporție. Estimarea I.C. pentru o proporție folosind SPSS presupune efectuarea unui set de operații.

1. Calculul *estimației proporției* unei categorii la nivelul eșantionul observat, $f = \frac{n_A}{n}$, unde n_A este numărul unităților din eșantion din categoria A, iar n volumul eșantionului. Acest calcul presupune demersul: meniul *Analyze* → comanda *Descriptive Statistics* → opțiunea *Frequencies*. În continuare se parcurg următorii pași:
   - în fereastra dialog *Frequencies* (vezi figura 7.12), selectăm variabila de interes (în exemplu nostru, variabila „Sexul persoanei") și o mutăm în zona *Variable*. Cerem *tabelul de frecvențe*, prin bifare în caseta de validare *Display frequency tables*;
   - prin butonul *OK*, se comandă obținerea output-ului.
2. Se află *valoarea variabilei Z* pentru nivelul de încredere considerat. De regulă, este folosit un nivel de încredere de 95%, căruia îi corespunde un Z = 1,96.
3. Se calculează *eroarea standard* (eroarea medie de selecție) S_p după relația: $S_p = \frac{s}{\sqrt{n}}$, unde $s = \sqrt{f(1-f)}$ este abaterea (deviația) standard, iar n este volumul eșantionului.
4. Se calculează limitele intervalului, folosind formula: $f \pm 1{,}96 S_p$.

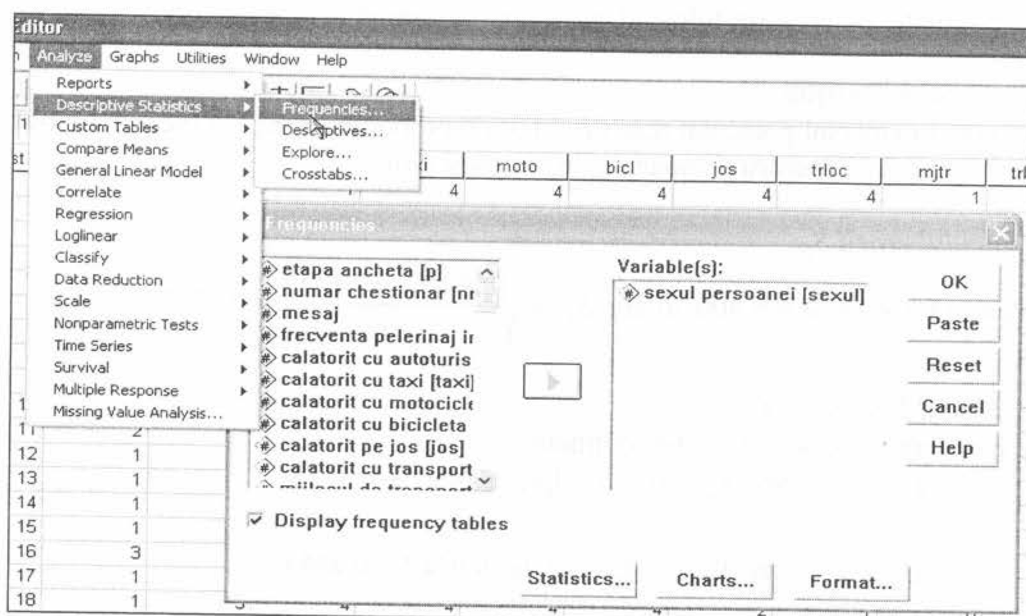


Figura 7.12 *Demersul pentru obținerea tabelului de frecvențe*

Frequencies

Statistics

sexul persoanei

N	Valid	400
	Missing	0

sexul persoanei

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	masculin	170	42,5	42,5	42,5
	feminin	230	57,5	57,5	100,0
	Total	400	100,0	100,0	

Figura 7.13 *Tabelul de frecvențe pentru variabila „Sexul persoanei", din* tapestry.sav

Exemplu. Dorim să estimăm prin interval de încredere proporția bărbaților în populația pelerinilor, folosind eșantionul prezentat în *tapestry.sav.*

Calculul proporției

Urmând demersul prezentat mai sus, aflăm o proporție de 42,5% (vezi output-ul din figura 7.13) pentru persoanele de sex masculin.

Calculul erorii S_p

Pentru f = 42,5%, n = 400, aflăm $S_p = \sqrt{\frac{0,425(1-0,425)}{400}}$ = 0,0247;

Calculul limitelor I.C.

Considerând un scor z = 1,96, obținem:

$$L_i = f - 1,96 \cdot S_p = 0.425 - 1,96 \cdot 0,0247 = 0,3765,$$

$$L_s = f + 1,96 \cdot S_p = 0.425 + 1,96 \cdot 0,0247 = 0,4895.$$

Interpretare. Ne putem aştepta, cu o încredere de 95%, ca procentul populaţiei de sex masculin în totalul pelerinilor la Iaşi, în 2002, să fie cuprins între 37,7% şi 48,9%. S-ar putea spune că, dacă s-ar repeta studiul de 100 de ori (adică s-ar înregistra 100 de eşantioane, independente şi identic observate), datele obţinute pentru 95 de eşantioane ar da acelaşi interval de încredere, numai 5 din cele 100 de eşantioane fiind susceptibile să dea valori în afara limitelor I.C. calculat.

CAPITOLUL 8

TESTAREA IPOTEZELOR STATISTICE

- **Demersul testării unei ipoteze statistice**
- **Teste parametrice în SPSS asupra mediilor şi proporţiilor**
- **Teste neparametrice în SPSS**

Testarea statistică este un procedeu prin care, în funcție de anumite reguli de decizie, se poate respinge sau nu o ipoteză formulată asupra unui parametru sau asupra unei distribuții.

8.1 Demersul testării unei ipoteze statistice

Demersul testării unei ipoteze presupune parcurgerea unor etape și rezolvarea problemelor pe care le implică, și anume:

1. Se formulează ipotezele, în funcție de problema pusă;
2. Se alege un test statistic în funcție de distribuția de selecție a statisticii considerate. Se alege un estimator $\hat{\theta}$ al parametrului θ de testat;
3. Se alege un prag de semnificație α pentru test;
4. Se stabilesc regulile de decizie, definind regiunile de „acceptare" și de „respingere" a ipotezei H_0;
5. Se calculează valoarea statisticii test, folosind datele înregistrate prin sondaj;
6. Se compară valoarea calculată a statisticii test cu valoarea teoretică;
7. Se ia decizia de a nu respinge sau de a respinge ipoteza admisă.

8.1.1 Ipoteze statistice

O ipoteză statistică este o presupunere cu privire la un parametru al unei distribuții date sau cu privire la legea de probabilitate a populației studiate.

Exemplu: ipoteza de egalitate a mediilor pentru a verifica dacă sunt diferențe semnificative între populațiile din care s-au extras eșantioanele observate.

În procesul de testare statistică, se formulează ipoteza nulă și ipoteza alternativă.

Ipoteza nulă (ipoteza de nul). Ipoteza nulă, ipoteza pe care dorim să o testăm, este notată H_0. Prin ipoteza nulă H_0 se admite, în principal, că nu există nici o diferență între valorile comparate. Ipoteza nulă H_0 este ipoteza pe care, de fapt, vrem să o discreditäm.

Ipoteza alternativă. Ipoteza alternativă, ipoteza pe care dorim să o testăm în opoziție cu ipoteza nulă, se notează cu H_1. Ipoteza alternativă este cea care va fi acceptată dacă, prin regula de decizie, se va respinge ipoteza nulă. Ipoteza H_1 este cea pe care, de fapt, vrem să o dovedim ca fiind adevărată.

Ipotezele asupra parametrului iau una din următoarele *trei forme*:

$$(1)\ \begin{matrix} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{matrix} \quad (2)\ \begin{matrix} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{matrix} \quad (3)\ \begin{matrix} H_0: \theta = \theta_0 \\ H_1: \theta > \theta_0 \end{matrix}$$

Prima formă a ipotezei alternative presupune un *test bilateral*, iar următoarele două un *test unilateral*.

Sunt considerate teste unilaterale şi testele de forma:

$$(2')\ \begin{matrix} H_0: \theta \geq \theta_0 \\ H_1: \theta < \theta_0 \end{matrix} \quad (3')\ \begin{matrix} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_0 \end{matrix}$$

Observaţie! Egalitatea apare întotdeauna în ipoteza nulă. Ipoteza alternativă se alege în funcţie de *ce vrem să probăm*: $\theta \neq \theta_0$, $\theta < \theta_0$, $\theta > \theta_0$.

8.1.2 Erori de testare

Prin definiţie, *eroarea este o diferenţă între o valoare adevărată şi o valoare observată*. În cazul testării unei ipoteze, se pot produce erori de acceptare sau de respingere pe nedrept a unei ipoteze, numite erori de primă speţă şi de a doua speţă sau erori de tip I şi erori de tip II[1].

Distribuţiile erorilor sunt distribuţii de probabilitate; fiecărui tip de eroare i se asociază o probabilitate de producere.

Eroarea de tip I. Eroarea de tip I comisă în testarea ipotezelor constă în *decizia de a respinge ipoteza nulă H_0 când în realitate aceasta este adevărată.* Probabilitatea asociată erorii de tip I este notată cu α şi este numită *prag de semnificaţie* sau *risc acceptat* în luarea deciziei că H_0 este falsă, α = P (*respinge H_0 când H_0 este adevărată*). În practică, α este cunoscut sub denumirea de *risc al vânzătorului.*

Eroarea de tip II. Eroarea de tip II este eroarea comisă în testarea ipotezelor *prin luarea deciziei de a accepta ipoteza nulă H_0 atunci când aceasta este falsă.* Probabilitatea asociată erorii de tip II este notată cu β şi reprezintă riscul

1. Introducerea distincţiei între eroarea de tip I şi eroarea de tip II este datorată lui J. Neyman şi E.S. Pearson (1938).

de a decide că H_0 este adevărată când H_0 este falsă, β = *P* (*acceptă* H_0 *când* H_0 *este falsă*). Riscul β este cunoscut sub denumirea de *risc al cumpărătorului*.

Tabelul 8.1 *Tipuri de erori și probabilitățile asociate acestora în testarea ipotezelor*

Realitate	Decizie		Suma probabilităților
	H_0	H_1	
H_0 adevărată	Decizie bună ($1-\alpha$)	Eroarea de tip I (α)	$(1-\alpha)+\alpha=1$
H_1 adevărată	Eroarea de tip II (β)	Decizie bună ($1-\beta$)	$\beta+(1-\beta)=1$

8.1.3 Regiunea de respingere și regiunea de acceptare a unei ipoteze

Regiunea de respingere. Regiunea de respingere este intervalul dintr-o distribuție de probabilitate a unei statistici considerate în care se respinge ipoteza nulă H_0, ipoteza H_1 fiind adevărată. Rezultă că o estimație calculată a estimatorului $\hat{\theta}$ al parametrului θ trebuie să fie semnificativ diferită, inferioară sau superioară valoarii ipotetice θ_0.

Există deci un *prag critic* de la care o estimație tinde să confirme ipoteza H_1 și să respingă ipoteza H_0. Pragul critic este definit plecând de la *eroarea de testare* de a respinge ipoteza H_0 când H_0 este adevărată. Acestui tip de eroare îi corespunde regiunea de respingere, numită și *regiune critică*, pentru care se asociază o probabilitate α.

Observație! În general, pentru α (în SPSS, valoarea *Sig.*, nivel de semnificație) se consideră o valoare cuprinsă între 0,01 și 0,1. Complementar, se definește regiunea de acceptare.

Interpretarea nivelului de semnificație în funcție de valoarea *Sig.*:
- dacă *Sig.* < *0,01* – atunci *statistica test* este semnificativă pentru 99% ;
- dacă *Sig.* < *0,05* – atunci *statistica test* este semnificativă pentru 95% ;
- dacă *Sig.* < *0,1* – atunci *statistica test* este semnificativă pentru 90% ;
- dacă *Sig.* > *0,1* – atunci *statistica test* nu este semnificativă.

Regiunea de acceptare. Regiunea de acceptare a unei ipoteze, numită şi *interval de încredere* (vezi figura 8.1), este un interval în care, pe baza unui test, nu se respinge ipoteza H_0. Regiunii de acceptare a ipotezei H_0 i se asociază o probabilitate $1-\alpha$, numită şi *coeficient de încredere.*

În testarea ipotezelor, *regiunea care se defineşte este regiunea de respingere a ipotezei* H_0, pentru un prag de semnificaţie α acceptat.

Testul bilateral. Într-un test bilateral, în legătură cu un parametru *θ*, ipotezele ce pot fi puse sunt:

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

În testul bilateral, regiunea de respingere a ipotezei H_0 corespunde unui interval, divizat în două subintervale, delimitate la un capăt de o valoare critică, prag critic, iar la celălalt capăt de infinit, şi anume:

($-\infty$; valoarea critică inferioară] şi [valoarea critică superioară; $+\infty$).

Valorile critice – inferioară (L_i) şi superioară (L_s) – sunt definite de relaţiile:

$$L_i = \mu_{\hat{\theta}} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}} \text{ şi } L_s = \mu_{\hat{\theta}} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}},$$

unde:

$\mu_{\hat{\theta}}$ este *media distribuţiei de selecţie* a unei statistici $\hat{\theta}$;

$\sigma_{\hat{\theta}}$ este *eroarea medie de selecţie* a statisticii $\hat{\theta}$;

α este pragul de semnificaţie al testului; fiind un test bilateral, se consideră $\alpha/2$.

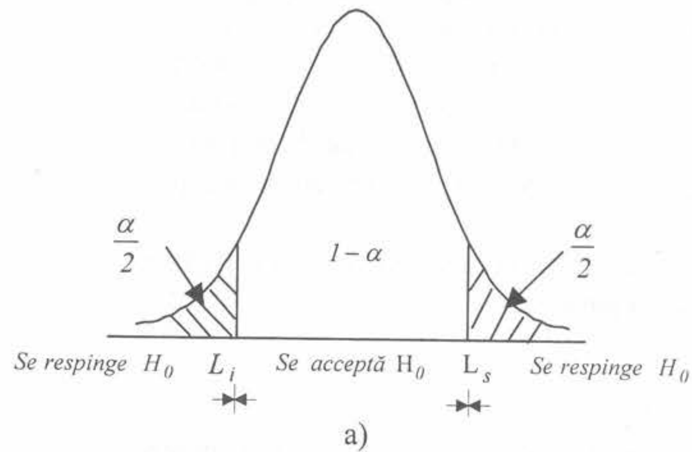
În figura 8.1.a se prezintă regiunea de acceptare şi regiunea de respingere a ipotezei H_0, în cazul testului bilateral.

Testul unilateral la dreapta. Pentru un test unilateral la dreapta, ipotezele sunt:

$$H_0: \theta = \theta_0$$

$$H_1: \theta > \theta_0$$

Într-un test unilateral la dreapta, regiunea de respingere a ipotezei H_0 este intervalul delimitat la stânga de valoarea critică $L_s = \mu_{\hat{\theta}} + z_\alpha \sigma_{\hat{\theta}}$.



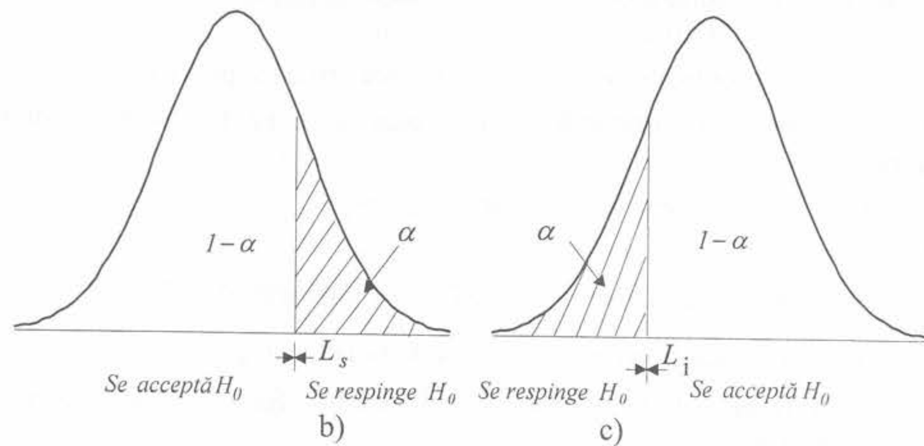


Figura 8.1 *Regiunea de acceptare și regiunea de respingere a ipotezei* H_0

Regiunea de respingere este egală cu mulțimea valorilor statisticii $\hat{\theta}$ cuprinse în intervalul [valoarea critică; ∞) și este reprezentată grafic în figura 8.1.b.

Testul unilateral la stânga. Pentru un test unilateral la stânga, ipotezele sunt:

$H_0: \theta = \theta_0$

$H_1: \theta < \theta_0$.

Într-un test unilateral la stânga, regiunea de respingere a ipotezei H_0 este intervalul delimitat la dreapta de valoarea critică $L_i = \mu_{\hat{\theta}} - z_\alpha \cdot \sigma_{\hat{\theta}}$.

Regiunea de respingere este egală cu mulţimea valorilor statisticii $\hat{\theta}$ cuprinse în intervalul $(-\infty$; valorea critică] şi este reprezentată grafic în figura 8.1.c.

8.1.4 Tipuri de teste

În funcţie de ipotezele formulate, de tipul variabilei/variabilelor considerate, de volumul populaţiei/populaţiilor şi de informaţiile disponibile asupra acestora, pot fi aplicate fie teste parametrice, fie teste neparametrice.

Teste parametrice. Aplicarea testelor parametrice presupune cunoaşterea formei parametrice a unei distribuţii a populaţiei considerate, adică a legii de distribuţie. Cel mai cunoscut test parametric este testul t – testul *Student*, propus de Gosset, în 1908. Acest test vizează compararea mediei unei populaţii (μ) cu o valoare fixă (μ_0) sau compararea mediilor a două populaţii care urmează o distribuţie normală. Este folosit, de asemenea, pentru testarea valorii unui coeficient de regresie, precum şi a valorii coeficientului de corelaţie. Alte teste foarte mult folosite sunt testele F şi χ^2.

Restricţii pentru aplicarea testelor parametrice. În procesul testării parametrice intervin mai multe elemente: un eşantion, o distribuţie de selecţie şi o populaţie şi, ca urmare, anumite ipoteze cu privire la parametri, care cer ca toate elementele considerate să fie compatibile unele cu altele. De exemplu, în ANOVA se pleacă de la ipotezele de independenţă, normalitate şi homoscedasticitate, adică:

- Observaţiile sunt independente;
- Datele sunt normal distribuite;
- Variabilele observate au aceeaşi varianţă.

Teste neparametrice. Testele neparametrice presupun testarea ipotezelor statistice fără a cere specificarea formei parametrice a distribuţiei populaţiilor comparate. Cele mai cunoscute teste neparametrice sunt: testul Wilcoxon (1945), folosit pentru a verifica, pe baza datelor de sondaj, dacă există diferenţe semnificative între două populaţii; testul Mann-Whitney (1947), folosit pentru verificarea existenţei egalităţii între două populaţii; testul Kolmogorov-Smirnov (1933), care vizează testarea identităţii a două funcţii de repartiţii (legi de distribuţie) etc.

8.2 Teste parametrice în SPSS asupra mediilor şi proporţiilor

Testele asupra mediilor, respectiv a proporţiilor, sunt folosite pentru a verifica dacă o medie/proporţie diferă semnificativ de o valoare specificată (ipotetică) sau pentru a compara două ori mai multe medii/proporţii între ele spre a testa dacă există diferenţe semnificative între ele (dacă eşantioanele observate provin din aceeaşi populaţie).

8.2.1 Alegerea testului

Testarea mediei cu o valoare specificată. Admitem ca parametru μ – nivelul mediu al distribuţiei unei populaţii – şi un estimator al acestuia, $\hat{\mu}$, respectiv o valoare $\bar{x}$ a estimatorului $\hat{\mu}$, care estimează valoarea parametrului μ.

În testarea ipotezelor cu privire la media unei populaţii, alegerea statisticii test depinde de volumul eşantionului (*n*) extras din populaţie şi de cunoaşterea sau nu a varianţei distribuţiei (σ^2, respectiv, s^2).

a) În *cazul eşantioanelor de volum mare* ($n \geq 30$), se foloseşte statistica test *Z*, care urmează o distribuţie de probabilitate normală, $Z \sim N(0,1)$.

– Când σ este *cunoscut*, statistica test *Z* este:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} .$$

– Când σ *nu este cunoscut*, statistica test *Z* este:

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} .$$

b) În *cazul eşantioanelor de volum mic* (n < 30) se foloseşte statistica test *t*, definită de relaţia:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} .$$

Statistica t urmează o distribuție de probabilitate *Student* cu $n - 1$ grade de libertate, $t \sim t(n - 1)$.

Testarea unei proporții. În cazul comparării unei proporții cu o valoare specificată, statistica test este:

$$\frac{f - p_0}{\sqrt{\frac{p(1-p)}{n}}}, \text{ respectiv } \frac{f - p_0}{\sqrt{\frac{f(1-f)}{n}}}.$$

Testarea a două medii. Admitem ca parametri μ_1 și μ_2 – nivelul mediu al distribuției pentru două populații, respectiv două valori $\overline{x}$ și $\overline{x}_2$ ale estimatorilor corespunzători, $\hat{\mu}_1$ și $\hat{\mu}_2$.

În *cazul eșantioanelor de volum mare* $(n \geq 30)$ se folosește statistica test Z, care urmează o distribuție de probabilitate normală, $Z \sim N(0,1)$.

– Când σ_1 și σ_2 sunt *cunoscute*, statistica test Z este:

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

– Când σ_1 și σ_2 sunt *necunoscute*, statistica test este:

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Pentru testarea a două proporții, statistica test este definită în mod analog statisticii test pentru două medii, și anume:

$$\frac{f_1 - f_2}{\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}}.$$

8.2.2 Testarea egalității unei medii cu o valoare specificată (One-Sample T Test și Error bar)

One-Sample T Test este un procedeu prin care se testează dacă media unei variabile este egală cu o constantă specificată (fie obținută în alt eșantion extras din aceeași populație, fie o valoare precizată, standard etc.)

În SPSS, testarea egalității unei medii cu o valoare specificată (valoare ipotetică) se poate realiza și printr-un procedeu grafic: *Error Bar*.

Testarea egalității unei medii cu o valoare specificată, folosind *One-Sample T Test*, presupune parcurgerea următorului demers: meniul *Analyze* → comanda *Compare Means* → opțiunea *One-Sample T Test* (vezi figura 8.2).

Exemplu. Considerând variabila „Vârsta persoanei" din *Tapestry.sav*, dorim să verificăm dacă vârsta persoanelor din eșantionul observat diferă semnificativ de valoarea 30 de ani.

După selectarea opțiunii *One-Sample T Test*, se parcurg următorii pași (vezi figura 8.3):

- Selectăm în fereastra *One-Sample T Test* variabila *vârsta* și o mutăm în zona *Test Variable(s)*;
- Specificăm valoarea dorită, *30*, în zona de editare *Test Value*;
- Activăm butonul de comandă *Options* care deschide fereastra *One--Sample T Test: Options* în care, în zona *Confidence Interval*, alegem gradul de încredere *95%*, după care acționăm butonul de comandă *Continue* pentru a reveni în fereastra *Sample T Test*;
- Acționăm butonul *OK* și comandăm SPSS obținerea output-ului.

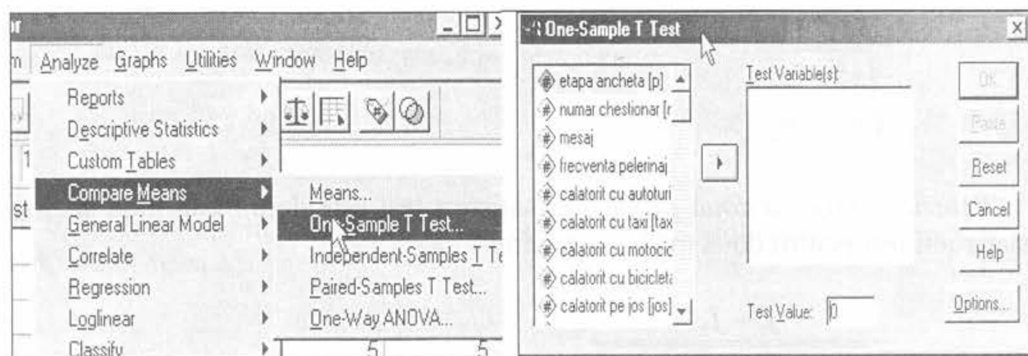
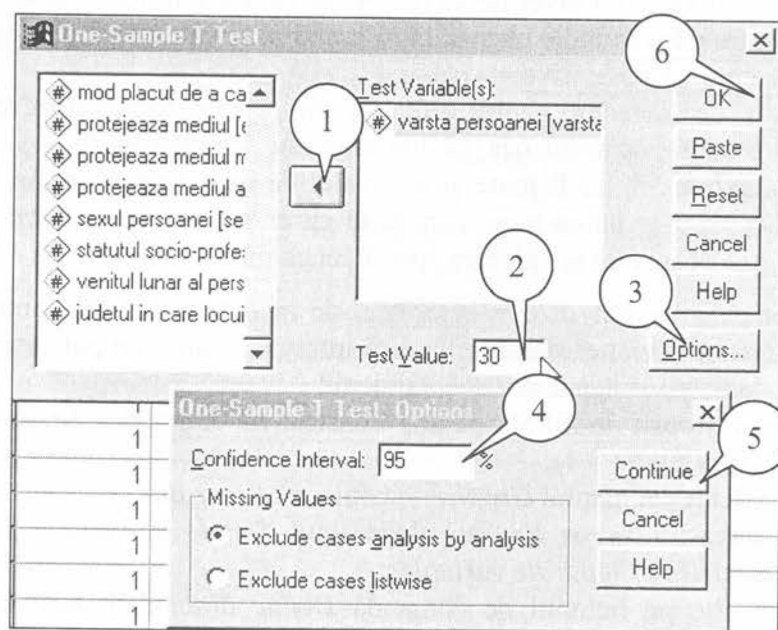


Figura 8.2 *Selectarea opțiunii* One-Sample T Test *și fereastra de dialog corespunzătoare*

Interpretare

Rezultate. Output-urile, *One-Sample Statistics* şi *One-Sample T Test* pentru variabila „Vârsta persoanei" (vezi figura 8.3) prezintă: valoarea medie observată egală cu *38,03* ani; valoarea specificată egală cu *30* ani; diferenţa dintre valoarea medie observată şi valoarea ipotetică de *8,03* ani.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
varsta persoanei	400	38.03	15.07	.75

One-Sample Test

	Test Value = 30					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
varsta persoanei	10.667	399	.000	8.03	6.55	9.52

Figura 8.3 *Comenzi în fereastra* One Sample T Test *şi output-ul corespunzător*

Statistica test. În exemplul luat, valoarea nivelului de semnificație *Sig.* (probabilitate) egală cu *0,000* este mai mică decât valoarea *0,05*, considerată în *Confidence Interval* din *Options*, ceea ce arată că există o *diferență semnificativă* între valoarea medie observată și cea specificată. În exemplul dat, deoarece valoarea *Sig.*< *0,01*, înseamnă că valoarea estimată a *statisticii test t* este semnificativă la un nivel de încredere de *99%*. Ca urmare, ipoteza nulă se respinge; între vârsta medie observată în eșantion și valoarea ipotetică (30 ani) există diferențe semnificative.

Intervalul de încredere pentru diferența dintre cele două valori nu conține zero, ceea ce arată, de asemenea, că diferența este semnificativă.

Dacă valoarea *Sig.* ar fi mare și intervalul de încredere ar conține valoarea zero, atunci nu s-ar putea trage concluzia că există o diferență semnificativă între valoarea observată și valoarea specificată a mediei.

Diagrama *Error Bar* descrie intervalul de încredere de 95% a mediei unei variabile (sau a deviației standard), adică intervalul care, am putea spune, cu o încredere de 95%, că acoperă valoarea medie.

Demersul folosit în SPSS pentru construirea diagramei *Error Bar* este următorul (vezi figura 8.4):

- Se selectează meniul *Graphs* → comanda *Error Bar*;
- În fereastra *Error Bar* se alege tipul *Simple* și butonul de opțiuni *Summaries of separate variables*;
- Prin clic pe butonul de comandă *Define* deschidem fereastra *Define Simple Error Bar*;
- Selectăm variabila considerată și o mutăm în zona *Error Bars*;
- În zona de editare *Level*, alegem intervalul de încredere pentru medie (implicit este 95%);
- Activăm butonul *OK* pentru a comanda obținerea diagramei dorite.

Pentru exemplificare considerăm aceleași date ca în procedeul *One-Sample T Test*.

Interpretare

Și prin procedeul *Error Bar* se poate observa că valoarea specificată (30) nu este cuprinsă în intervalul de încredere (36,5; 39,5). Ca urmare, se poate spune cu o *încredere de 95%* că se respinge ipoteza de nul, adică de egalitate a vârstei medii a persoanelor din eșantion cu valoarea specificată (30).

Observație! Se constată că prin cele două procedee, *One-Sample T Test* și *Error Bar*, s-a ajuns la aceeași concluzie: respingerea ipotezei de nul; între vârsta medie observată la nivelul eșantionului considerat și valoarea specificată există o diferență semnificativă.

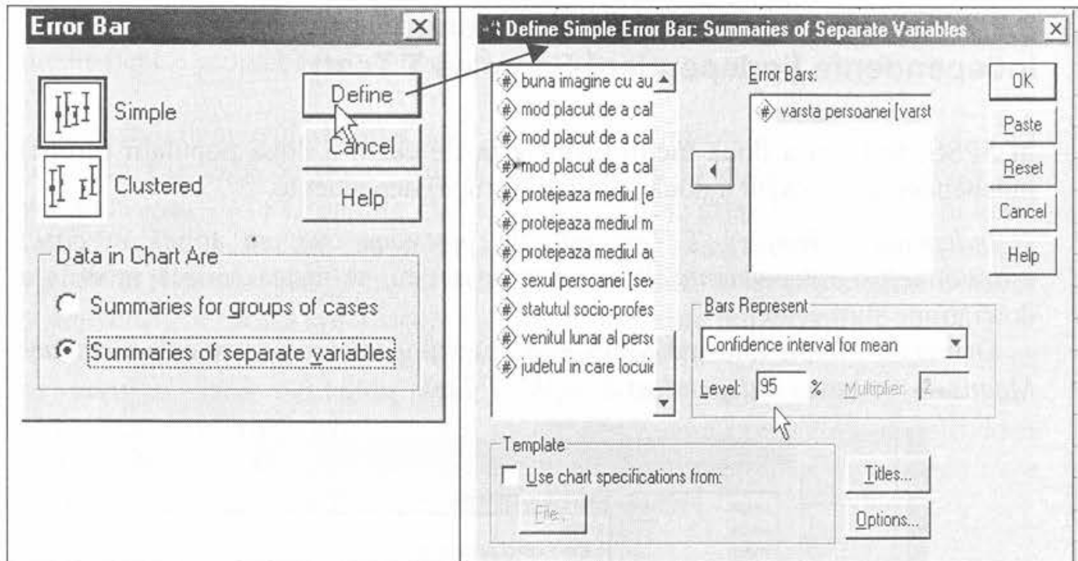


Figura 8.4 *Ferestrele* Error Bar *şi* Define Simple Error Bar

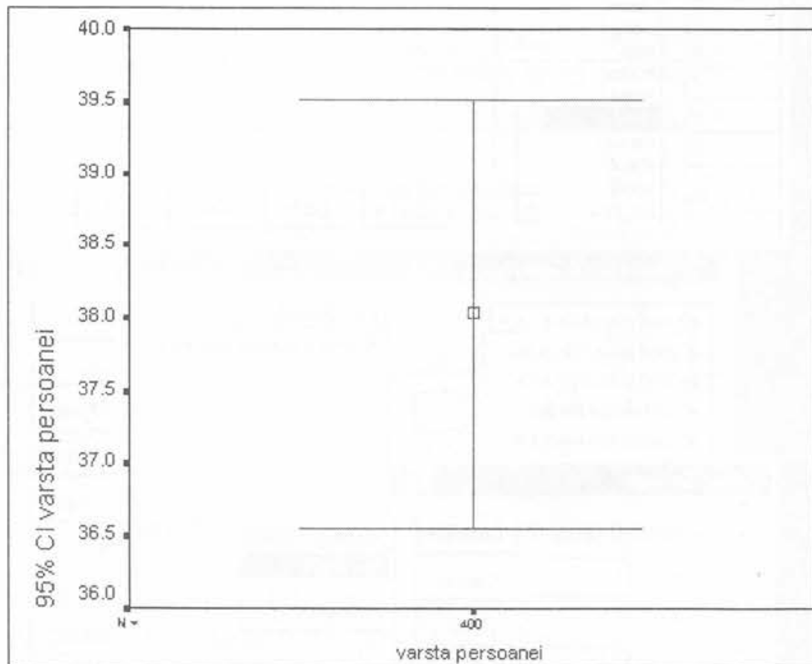


Figura 8.5 *Diagrama* Error Bar – *Intervalul de încredere 95% pentru media variabilei „Vârsta"*

8.2.3 Testarea egalității mediilor a două eşantioane independente (Independent-Samples T Test)

În SPSS, testarea a două medii poate viza fie cazul a două populații (grupe) independente, fie cazul a două populații (grupe) dependente.

Independent-Samples T Test este un procedeu care se aplică în *cazul eşantioanelor independente.* Prin acest procedeu, se testează dacă mediile a două grupe sunt egale.

Demersul testării folosind SPSS este: meniul *Analyze* → comanda *Compare Means* → opțiunea *Independent-Samples T Test.*

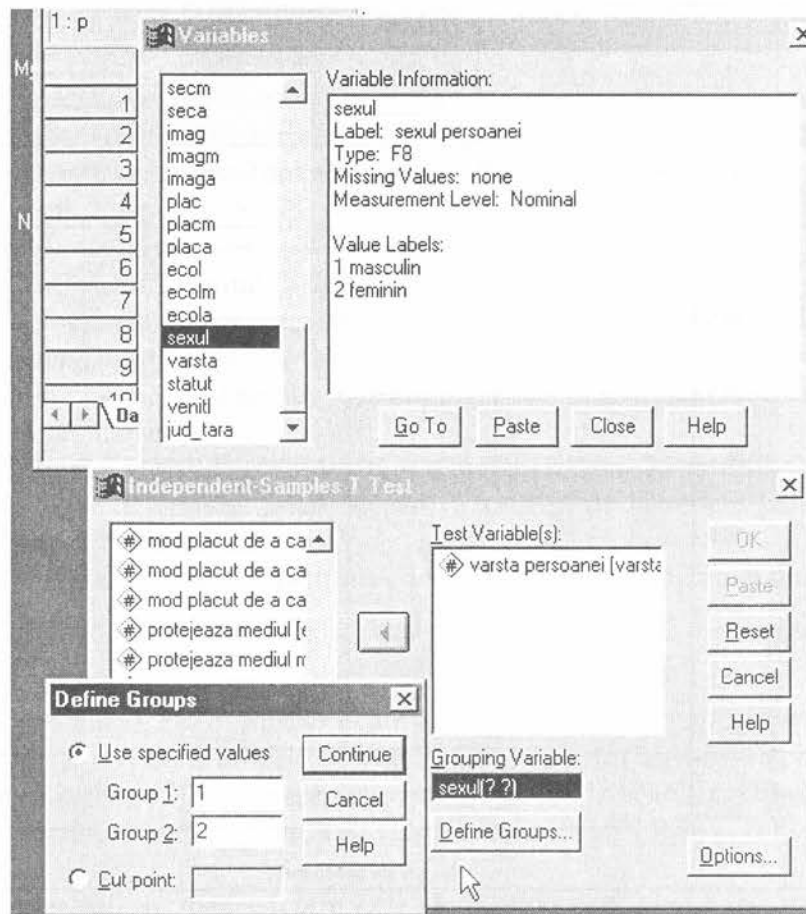


Figura 8.6 *Fereastra* Variables (*meniul* Utilities → *comanda* Variables) *şi fereastra* Independent-Samples T Test (*meniul* Analyze → *comanda* Compare Means)

Exemplu: Dorim să testăm dacă, la nivelul eşantionului observat, vârsta medie pentru grupa bărbaţi este diferită de vârsta medie pentru femei.

Urmând demersul de mai sus, în fereastra *Independent-Samples T Test* selectăm variabila de testat „vârsta" şi o mutăm în zona *Test Variables*, iar variabila „sex" – în *Grouping Variable*. Se definesc grupele variabilei; în acest caz, folosim 1 pentru masculin şi 2 pentru feminin. Informaţiile asupra variabilei le putem găsi şi în fereastra *Variables* dacă selectăm meniul *Utilities* şi comanda *Variables* (vezi figura 8.6). Output-ul este prezentat în figura 8.7.

Calculul *statisticii test* pentru compararea mediilor a două populaţii cere să se verifice dacă deviaţiile standard la nivelul celor două grupe sunt semnificativ diferite, deoarece prin ipoteza de nul se presupune că cele două populaţii au varianţe egale. Se foloseşte în acest scop *testul Levene de egalitate a varianţelor* (*Levene's Test for equality of Variance*).

Group Statistics

	sexul persoanei	N	Mean	Std. Deviation	Std. Error Mean
varsta persoanei	masculin	170	37.35	14.90	1.14
	feminin	230	38.54	15.20	1.00

Independent Samples Test

		Levene's Test for quality of Variance		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed	Mean Difference	Std. Error Difference	Lower	Upper
varsta persoa	Equal varianc assumed	.168	.682	-.785	398	,433	-1.20	1.52	-4.19	1.80
	Equal varianc not assumed			-.787	368.299	.432	-1.20	1.52	-4.19	1.79

Figura 8.7 *Output-ul din* Independent-Samples T Test

Interpretare. Dacă nivelul de semnificaţie observat pentru acest test este mic (de exemplu, mai mic decât 0,05), atunci se folosesc *varianţe distincte* pentru testarea mediilor. Dacă acest nivel este mare, ca în cazul considerat (*Sig.* este egal cu 0,682), atunci se folosesc *varianţe reunite* sub formă de medie ponderată s_p^2.

Test cu varianțe comune:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Test cu varianțe separate:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

unde:

$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ reprezintă varianța comună;

$\bar{x}_i$ – media grupei *i*;

n_i – numărul observațiilor din grupa *i*;

s_i^2 – varianța grupei *i* la nivelul eșantionului observat;

($n_1 + n_2 - 2$) reprezintă gradele de libertate pentru testul cu varianțe comune.

Nivelul de semnificație (*Sig.*) pentru testul *Levene* fiind mare (0,682), în exemplul dat, folosim testul cu varianțe comune (*Equal variances assumed*).

În acest caz, *testul t* este egal cu – 0,785, cu 398 grade de libertate și o probabilitate *Sig.* de 0,433 (mai mare decât 0,05), și ne arată că pentru mediile celor două grupe (37,35 și 38,54) nu se poate trage concluzia că diferă semnificativ.

La aceeași constatare ajungem și prin observarea intervalului de încredere pentru diferența dintre cele două valori. Intervalul conține zero, ca urmare nu se poate trage concluzia că diferența dintre valorile medii ale celor două grupe este semnificativă.

8.2.4 Testarea egalității mediilor a două eșantioane perechi (Paired-Samples T Test)

Paired-Samples T Test este un procedeu care se aplică în *cazul eșantioanelor dependente*. Prin acest procedeu, se compară mediile pentru un singur grup observat în momente diferite.

Adesea, prin acest test se observă aceiași subiecți în două momente diferite, de exemplu, înainte și după un tratament, verificându-se dacă diferențele dintre valorile medii sunt semnificative. Se calculează *diferențele dintre valorile celor două variabile pentru fiecare caz în parte* și se testează dacă diferențele dintre mediile acestora diferă de zero.

Demersul folosit în SPSS este: meniul *Analyze* → comanda *Compare Means* → opțiunea *Paired-Samples T Test*.

Exemplu: Considerăm variabilele *p_inv00* şi *p_inv01* din *dez_reg.sav*. Dorim să verificăm dacă nivelul mediu al numărului populaţiei ocupate în învăţământ la nivel de judeţ, în România, în anul 2000, diferă sau este echivalent cu cel din anul 2001 (la sfârşitul anului).

Demersul urmat, după selectarea opţiunii *Paired-Samples T Test*, este prezentat în figura 8.8.

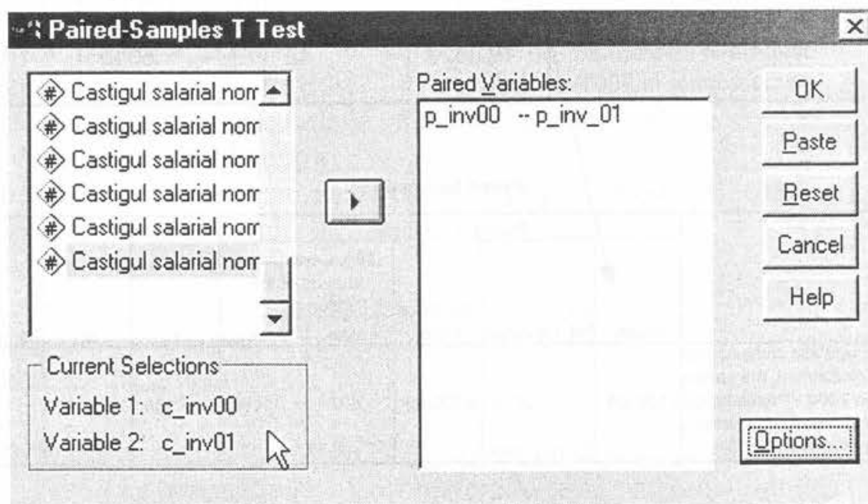


Figura 8.8 *Fereastra* Paired-Samples T Test

- Selectăm în fereastra dialog *Paired-Samples T Test* prima variabilă, prin clic asupra ei, şi vom vedea că SPSS o mută în *Current Selections* (în partea din stânga jos a ferestrei), ca *Variable 1*;
- Efectuăm aceeaşi operaţiune pentru a doua variabilă;
- Mutăm perechea de variabile în zona *Paired Variables* (în partea dreaptă a ferestrei dialog). Se pot repeta aceste operaţii pentru câte perechi de interes avem;
- Prin butonul de comandă *OK*, se obţine output-ul prezentat în figura 8.9.

Interpretare. Pentru *testul t*, corespunzător procedeului *Paired Samples Test*, s-au calculat, mai întâi, pentru fiecare judeţ în parte, diferenţele dintre valorile din 2000 şi 2001, apoi s-au calculat *media acestor diferenţe* ($\overline{x}_d$) şi *deviaţia standard a diferenţelor* (σ_d). Pe baza acestor rezultate se obţine *valoarea statisticii test t*:

$$t = \frac{\overline{x}_d}{\sigma_d / \sqrt{n}}.$$

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Populatia civila ocupata in invatamant, mii persoane, in 2000	10.0333	42	7.4737	1.1532
	Populatia civila ocupata in invatamant, mii persoane, in 2001	10.0429	42	7.5310	1.1621

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Populatia civila ocupat invatamant, mii persoa in 2000 - Populatia civ ocupata in invatamant persoane, in 2001	9.52E-03	.3773	.822E-02	-.1271	.1080	-.164	41	.871

Figura 8.9. *Output-ul* Paired Samples Test

Media diferențelor perechi între populația ocupată în învățământ în România în anul 2000 și cea din 2001 este de 96 mii de persoane. Valoarea *Sig.* (egală cu 0,871) asociată cu statistica *test t* este mare (> 0,05), fapt ce nu ne permite să concluzionăm că media diferențelor perechi de 0,00952 (9,52E – 03) este diferită semnificativ de zero.

8.2.5 Testarea egalității a trei și mai multe medii (One-Way ANOVA)

ANOVA (*Analysis of Variance*) este un procedeu de analiză a varianței unei variabile numerice sub influența unei variabile de grupare.

Prin ANOVA, se compară medii pentru trei și mai multe subpopulații definite de variabila de grupare (variabila independentă).

Această metodă permite extensia analizei realizate prin *testul t*, aplicabil asupra a două medii, la situații în care variabila independentă (variabila de grupare) prezintă trei și mai multe categorii (niveluri).

De asemenea, ANOVA poate fi folosită în analiza unor situații în care asupra variabilei numerice (variabila dependentă) acționează simultan mai multe variabile independente. În astfel de cazuri, prin ANOVA se poate prezenta modul în care aceste variabile independente interacționează una cu alta şi ce efecte au aceste interacțiuni asupra variabilei dependente.

One-Way ANOVA (ANOVA unifactorială) este unul din procedeele de analiză a varianței pentru *o variabilă cantitativă* dependentă de *o singură variabilă factor* (de grupare). Variabila factor, numită şi variabilă independentă, explicativă, trebuie să fie calitativă şi trebuie să aibă un număr redus de categorii (modalități).

Ipoteza nulă, ipoteza de testat, formulată prin acest procedeu, presupune *egalitatea a trei şi mai multe medii*:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k,$$

unde:

μ_i este media grupei *i*.

Interpretarea rezultatelor ANOVA vizează două teste, şi anume:

- *Testul de omogenitate a varianțelor*. Această problemă implică testul de omogenitate a varianțelor subpopulațiilor, definite de modalitățile variabilei factor (de grupare). Acest test este necesar pentru a determina care test este adecvat comparării mediilor. Ipoteza de nul este respinsă dacă valoarea *Sig.* (probabilitatea α) este inferioară valorii 0,05 (5%), semnificând că nu sunt egale toate varianțele.
- *Testul ANOVA*. Ipoteza nulă este respinsă dacă valoarea *Sig.* este inferioară valorii 0,05 (5%), semnificând că cel puțin două medii, calculate la nivelul subpopulațiilor, diferă între ele.

În SPSS, pentru compararea a trei şi mai multe medii este folosit următorul demers: meniul *Analyze* → comanda *Compare Means* → opțiunea *One-Way ANOVA*.

Exemplu: Considerăm variabilele *p_inv01* şi *regiunea*. Dorim să verificăm dacă nivelul mediu al numărului populației ocupate în învățământ pe un județ, în România, în anul 2001, este acelaşi sau diferă de la o regiune la alta, respectiv dacă diferența dintre mediile grupelor (regiunile României) este egală cu zero sau este semnificativ diferită de zero.

După selectarea opțiunii *One-Way ANOVA*, se parcurg următorii paşi:

- În fereastra de dialog *One-Way ANOVA* alegem variabila *p_inv01* pe

care o mutăm în zona *Dependent List* şi variabila *regiunea* pe care o mutăm în zona *Factor*;

- Prin butonul de comandă *Options* (vezi figura 8.10), se deschide fereastra *One-Way ANOVA: Options* în care se bifează casetele de validare *Descriptive*, *Homogeneity of variance* şi *Means plot* pentru a se verifica îndeplinirea restricţiilor de *normalitate*, *homoscedasticitate* şi *independenţă* impuse unei analize ANOVA.

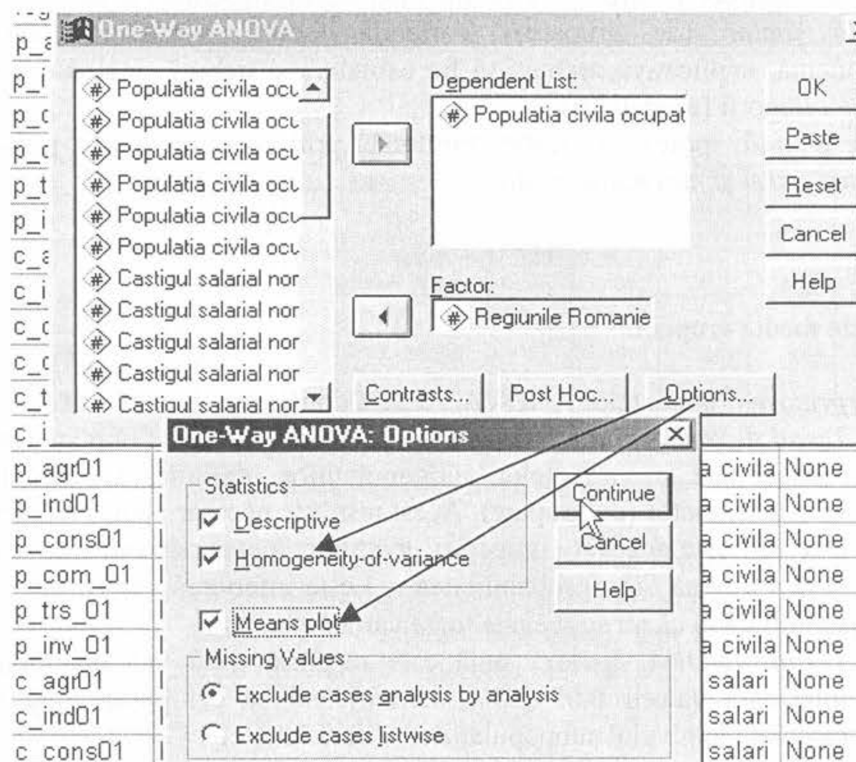


Figura 8.10 *Alegerea variabilelor şi statisticilor în procedeul* One-Way ANOVA

Restricţia de normalitate se verifică observând dacă distribuţia valorilor din fiecare grupă prezintă asimetrie accentuată, dacă sunt outlier-i sau alte anomalii. În acest scop, se pot folosi rezultatele din *Descriptives*. De asemenea, se pot utiliza diagramele *Boxplot*, create prin opţiunea *Explore* a comenzii *Descriptive Statistics* din meniul *Analyze* (vezi *Boxplot* din figura 8.11), sau alte procedee de verificare a normalităţii (vezi paragraful 6.4).

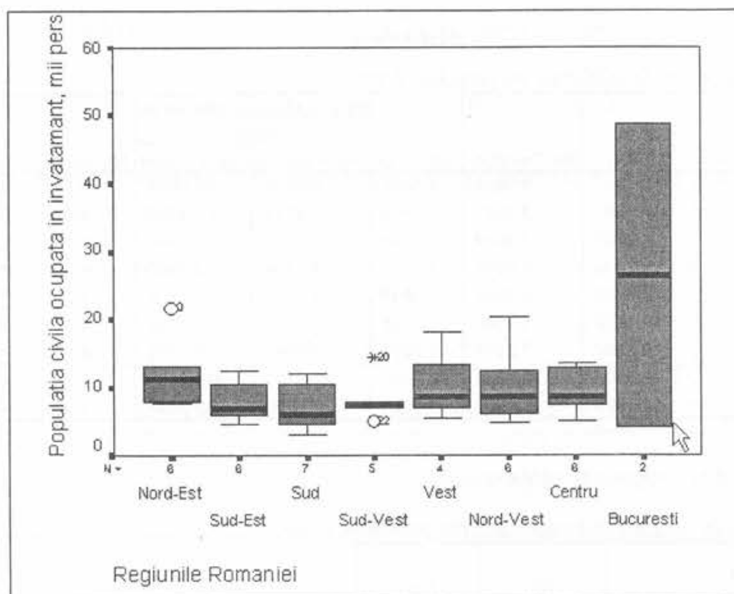


Figura 8.11 Boxplot

Observând diagramele *Boxplot*, se constată că cele 8 grupe au distribuţii normale după numărul persoanelor ocupate în învăţământ; prezintă o asimetrie relativă. De asemenea, se constată valori extreme (outlier, regiunea Bucureşti).

Restricţia de homoscedasticitate. Una din restricţiile aplicării ANOVA o constituie homoscedasticitatea, adică se presupune că varianţele grupelor sunt egale. Se poate verifica această ipoteză cu ajutorul testului *Levene – Test of Homogeneity of Variances.*

Output-ul pentru acest test este prezentat în figura 8.12.

Interpretare. Valoarea *Sig.* (testul Levene) egală cu 0,000 este mai mică decât 0,05, sugerând că varianţele pentru cele 8 regiuni nu sunt egale. În aceste condiţii, fiind încălcată *restricţia de homoscedasticitate*, nu se poate aplica ANOVA.

Observaţie! Se observă, în output-ul *Descriptives*, că regiunea Bucureşti are o valoare a deviaţiei standard care se abate mult de la valoarea celorlalte regiuni. Acest fapt ne motivează să considerăm regiunea Bucureşti ca outlier şi, ca urmare, să o excludem din ansamblul regiunilor. Pentru aceasta, din foaia *Variable View* a fişierului *Data Editor: Pop_ocupata pe judete.sav* selectăm coloana *Missing* pentru a deschide fereastra *Missing Values.* În această fereastră, selectăm butonul de opţiuni *Discrete missing values*, precizăm valorile outlier-e şi refacem aplicaţia în noile condiţii (fără această regiune). Noile rezultate sunt prezentate în figura 8.13.

Descriptives

Populatia civila ocupata in invatamant, mii persoane, in 2001

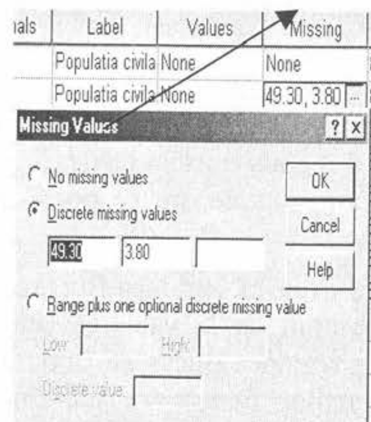
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Nord-Est	6	12.0667	5.3616	2.1889	6.4400	17.6933	7.70	21.80
Sud-Est	6	7.9667	3.2327	1.3198	4.5741	11.3592	4.40	12.80
Sud	7	7.2857	3.9019	1.4748	3.6771	10.8944	3.00	12.20
Sud-Vest	5	8.3400	3.6295	1.6231	3.8334	12.8466	4.90	14.50
Vest	4	9.6750	4.7528	2.3764	2.1122	17.2378	5.50	16.50
Nord-Vest	6	10.1500	5.7386	2.3428	4.1278	16.1722	4.70	20.10
Centru	6	9.3667	3.2042	1.3081	6.0041	12.7292	4.90	13.40
Bucuresti	2	26.5500	32.1734	22.7500	-262.5162	315.6162	3.80	49.30
Total	42	10.0429	7.5310	1.1621	7.6960	12.3897	3.00	49.30

Test of Homogeneity of Variances

Populatia civila ocupata in invatamant, mii persoane, in 2001

Levene Statistic	df1	df2	Sig.
19.569	7	34	.000

Figura 8.12 *Mediile, deviațiile standard calculate pe regiuni, precum și testul Levene de omogenitate a varianțelor*



Populatia civila ocupata in invatamant, mii persoane, in 20

	N	Mean	Std. Deviation
Nord-Est	6	12.0667	5.3616
Sud-Est	6	7.9667	3.2327
Sud	7	7.2857	3.9019
Sud-Vest	5	8.3400	3.6295
Vest	4	9.6750	4.7528
Nord-Vest	6	10.1500	5.7386
Centru	6	9.3667	3.2042
Total	40	9.2175	4.2908

Test of Homogeneity of Variances

Populatia civila ocupata in invatamant, mii persoane, in 20

Levene Statistic	df1	df2	Sig.
.572	6	33	.749

Figura 8.13 *Rezultatele după excluderea valorilor outlier ale regiunii București*

În noile condiții, valoarea *Sig.* (0.749) pentru testul de omogenitate a varianțelor este mai mare ca 0,05, sugerînd că varianțele pentru cele 7 regiuni sunt egale, deci restricția de homoscedasticitate este îndeplinită și astfel se poate aplica ANOVA.

Tabelul ANOVA și graficul corespunzător pentru mediile pe regiuni sunt prezentate în figura 8.14, respectiv figura 8.15.

ANOVA

Populatia civila ocupata in invatamant, mii persoane, in 2001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	94.255	6	15.709	.831	.555
Within Groups	623.783	33	18.903		
Total	718.038	39			

Figura 8.14 *Tabelul* ANOVA

Interpretare. În tabelul ANOVA din figura 8.14 sunt prezentate: *statistica test F*, *valoarea Sig.*, precum și elementele de calcul pentru *statistica test F.*

Statistica test F se calculează după relația:

$$F = \frac{S_E^2}{S_R^2},$$

unde:

S_E^2 reprezintă estimatorul *varianței intergrupe* (*Between-Groups*). Se calculează ca medie a pătratelor abaterilor mediei fiecărei grupe față de media pe ansamblul grupelor și arată varianța datorată influenței factorului de grupare;

S_R^2 reprezintă estimatorul *mediei varianțelor de grupă* și arată *varianța din interiorul fiecărei grupe* (*Within Groups*), varianța datorată influențelor aleatorii.

Cu cât mediile grupelor au valori mai diferite între ele, cu atât variația dintre grupe este mai mare; cu cât o variație, în interiorul grupelor, este relativ mai mică, cu atât statistica test F este mai mare, arătând că ipoteza nulă poate fi respinsă.

În exemplul considerat, *statistica test F* este mică (*0,831*), cu o probabilitate asociată *Sig. (0,555)* mai mare decât *0,05* – evidențiază că ipoteza de egalitate a mediilor pe grupe nu se respinge, deci regiunile nu diferă semnificativ în raport

cu numărul mediu al persoanelor ocupate în învăţământ, la nivelul anului 2001. Acest fapt se poate observa şi din figura 8.15.

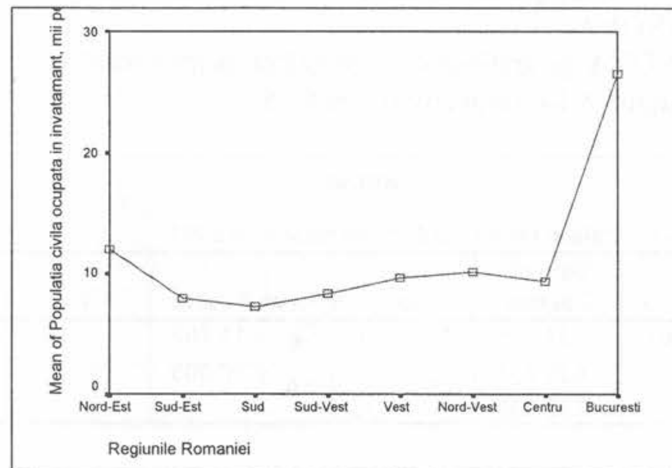


Figura 8.15 *Numărul mediu al populaţiei ocupate în învăţământ, pe regiuni, în România, 2001*

8.3 Teste neparametrice în SPSS

Testele neparametrice aplicabile în SPSS sunt: Chi-Square, Binomial, Runs, 1 Sample K-S, 2 Independent Samples, K Independent Samples, 2 Related Samples, K Related Samples (vezi figura 8.16).

8.3.1 Testarea egalităţii unei proporţii cu o valoare specificată (Binomial Test)

Binomial Test este un procedeu prin care se testează ipoteze cu privire la o variabilă cu distribuţie binomială, variabilă care poate lua doar două valori, de exemplu, sexul persoanelor.

Pentru astfel de variabile, se calculează frecvenţele de apariţie a fiecăreia dintre cele două valori, iar pe baza lor, media, deviaţia standard etc.

Binomial Test este similar cu *One Sample t-test* şi este folosit pentru a *compara o proporţie cu o valoare specificată*.

Realizarea acestui test în SPSS presupune următorul demers: meniul *Analyze* → comanda *Nonparametric Tests* → opțiunea *Binomial*.

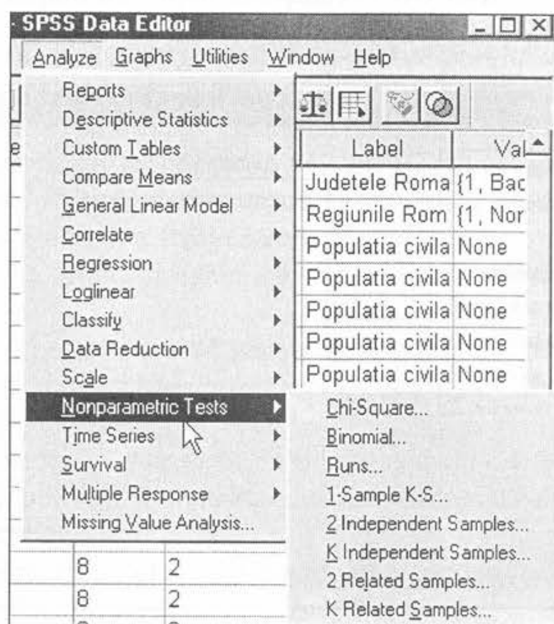


Figura 8.16 *Testele neparametrice aplicabile în SPSS*

Exemplu: Dorim să verificăm dacă proporția uneia dintre cele două grupe de persoane definite prin variabila „sexul persoanei”, masculin și feminin, diferă semnificativ de *0,50*. Suma proporțiilor, respectiv a probabilităților de apariție a celor două valori fiind unu, probabilitatea pentru oricare valoare este *0,50*, adică *1 minus 0,50*.

După selectarea opțiunii *Binomial* și deschiderea ferestrei *Binomial Test*, pașii de urmat sunt:

- În fereastra *Binomial Test* selectăm variabila binomială, „sexul persoanei” și o mutăm în zona *Test Variable List* (se pot selecta mai multe variabile);
- În zona *Define Dichotomy* alegem *Get from date*, adică cele două valori ale variabilei (1 pentru masculin și 2 pentru feminin) sunt definite în foaia *Variable View*. Cealaltă opțiune, *Cut point*, se folosește atunci când se consideră o variabilă continuă pe care o dichotomizăm. De exemplu, pentru variabila „vârsta persoanei”, am putea lua două grupe: 1 – grupa persoanelor cu vârsta $\leq$ 20 ani – și 2 – grupa persoanelor cu vârstă > 20 ani;

- În zona de editare *Test Proportion* se precizează valoarea dorită. Implicit, se consideră valoarea 0,50;
- Butonul de comandă *OK* declanşează obţinerea output-ului (vezi figura 8.18).

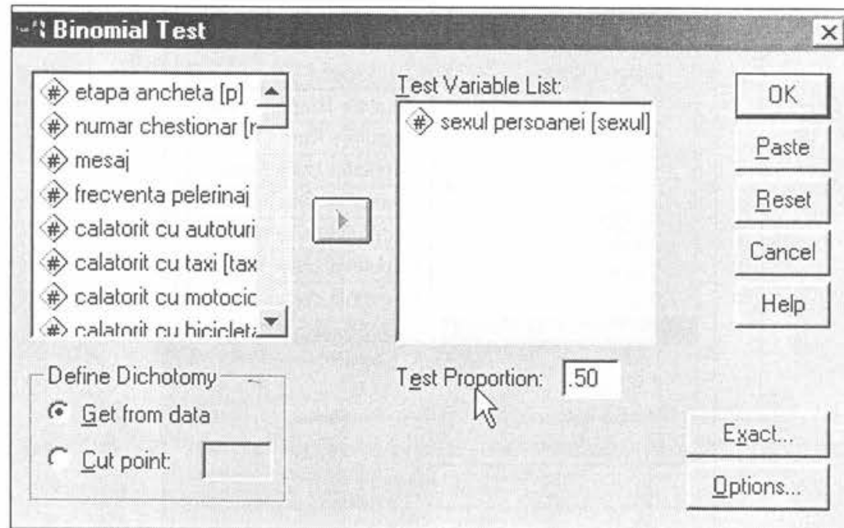


Figura 8.17 *Fereastra* Binomial Test

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
sexul persoane	Group 1	masculin	170	.43	.50	.003[a]
	Group 2	feminin	230	.57		
	Total		400	1.00		

a. Based on Z Approximation.

Figura 8.18 *Output-ul pentru* Binomial Test

Interpretare. Proporţia observată în eşantion pentru grupa 1 (masculin) este de 43%, proporţia specificată este 50%. Valoarea *Sig.* asociată testului este mai mică decât 0,05, astfel încât se poate concluziona, cu o încredere de 95%, că proporţia bărbaţilor în eşantion diferă semnificativ de proporţia specificată, 50%.

8.3.2 Testarea egalității a două şi mai multe proporții (Chi-Square)

În SPSS, *procedeul Hi-pătrat* (numit şi *Hi-pătrat de ajustare*) se aplică pentru testarea ipotezelor cu privire la variabile nominale (categoriale) sau variabile ordinale, fie ca test de „ajustare", fie ca test de independență.

În cazul unei variabile nominale, testul Hi-pătrat este folosit pentru a verifica dacă distribuția de frecvență a unei variabile pe categorii corespunde fie cu distribuția teoretică a frecvențelor relative (ipoteza de nul presupune că toate categoriile au proporții egale), fie cu o distribuție de frecvență propusă (rețetă).

Aplicarea acestui procedeu de testare presupune următorul demers: meniul *Analyze* → comanda *Nonparametric Tests* → opțiunea *Chi-Square Test.*

Exemplu: Considerăm variabila „sexul persoanei" din *Tapestry.sav*. Dorim să verificăm dacă proporția persoanelor de sex masculin este egală cu proporția persoanelor de sex feminin.

După selectarea opțiunii *Chi-Square Test* şi deschiderea ferestrei *Chi-Square Test*, paşii de urmat sunt:

- În fereastra dialog *Chi-Square Test* (vezi figura 8.19) selectăm variabila pentru care dorim să testăm proporțiile, în cazul nostru variabila „sexul persoanei" şi o mutăm în zona *Test Variable List*. Se pot selecta mai multe variabile, pentru fiecare variabilă obținându-se câte un tabel de frecvență separat;
- În zona *Expected Range* definim categoriile pentru care dorim să testăm proporțiile. Alegem *Get from data*, considerând categoriile definite pentru variabila „sexul persoanei".

Observație! În cazul când se lucrează cu o variabilă continuă, se alege *Use specified range*, specificându-se valoarea minimă şi valoarea maximă între care dorim să verificăm dacă elementele au aceeaşi pondere cu valoarea specificată.

- În zona *Expected Values*, alegem ipoteza *toate proporțiile egale* sau *proporții specificate* (rețetă). Optăm pentru *All categories equal*;
- Selectând butonul de comandă *Options* se deschide fereastra *Chi-Square Test: Options* în care se poate opta pentru *Descriptive* (media, deviația standard, valoarea minimă, valoarea maximă, numărul cazurilor) sau/şi pentru *Quantiles* (centila a 25-a, a 50-a, a 75-a).
- Prin clic pe butonul de comandă *Continue*, se revine în fereastra *Chi-Square Test*, din care se selectează *OK*, care comandă lansarea procedurii de obținere a output-ului, prezentat în figura 8.20.

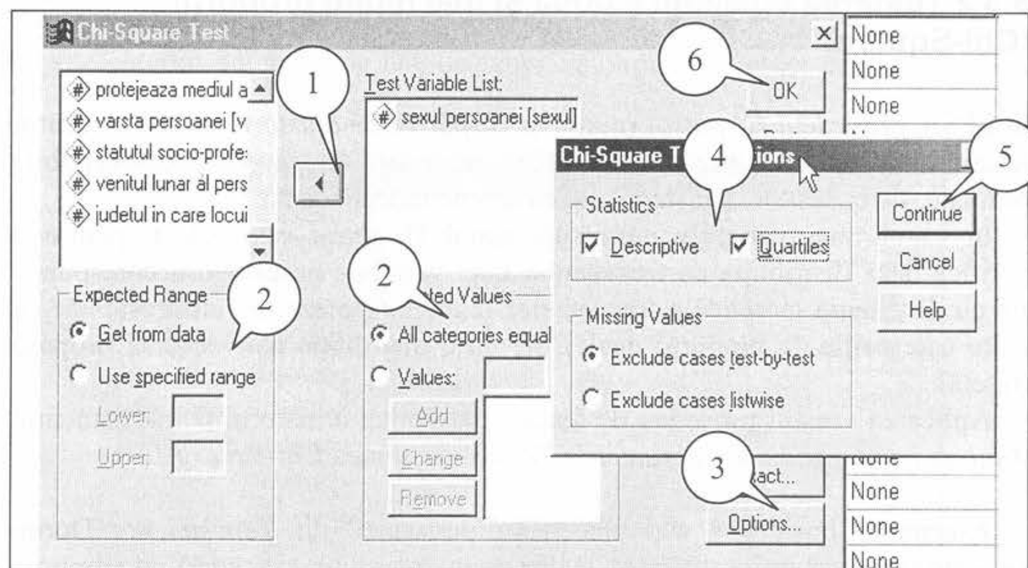


Figura 8.19 *Alegerea opțiunilor în fereastra* Chi-Square Test

Interpretare. În tabelul frecvențelor, sunt comparate *frecvențele observate* cu *frecvențele teoretice* (așteptate conform ipotezei de nul), pentru fiecare categorie *i*. *Diferențele* sunt prezentate pe categorii în coloana *Residual*. În exemplu dat, se observă că sunt 170 de persoane de sex masculin și 230 de sex feminin. Conform ipotezei de nul (de egalitate a proporțiilor), pentru fiecare categorie ar trebui să fie câte 200 de persoane. În coloana *Residual* sunt prezentate diferențele față de valorile teoretice, pentru fiecare categorie, și anume: – 30 și 30.

În tabelul *Chi Square Test*, se prezintă valoarea statisticii *Hi-pătrat* (*Chi--Square* – χ^2), gradele de libertate (*df*) și valoarea semnificației (*Asymp. Sig.*).

sexul persoanei

	Observed N	Expected N	Residual
masculin	170	200.0	-30.0
feminin	230	200.0	30.0
Total	400		

Test Statistics

	sexul persoanei
Chi-Square[a]	9.000
df	1
Asymp. Sig.	.003

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 200.0.

Figura 8.20 *Output-ul pentru procedeul* Hi-pătrat *în cazul unei variabile categoriale*

Statistica test Hi-pătrat pentru o variabilă se calculează după relația:

$$\chi^2 = \sum_i \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

unde:

n_i reprezintă frecvențele observate în categoria *i*;

n = $\sum_i n_i$ reprezintă volumul eşantionului;

p_i reprezintă frecvența relativă teoretică, $\sum_i p_i = 1$. Fiecare p_i se înmulțeşte cu *n* pentru a deveni comparabilă cu frecvența observată, n_i.

În exemplul dat, valoarea estimată a statisticii *Hi-pătrat* este semnificativă la un nivel de încredere de 99%, deoarece valoarea *Asimp.Sig.< 0,01*. Ca urmare, ipoteza nulă este respinsă. Se poate trage concluzia că cele două categorii de persoane (masculin, feminin) nu au aceeaşi proporție; distribuția nu este uniformă.

[illegible]

[illegible]

[illegible] $\sum$ [illegible]

[illegible] eigenvalues [illegible]

[illegible] eigenvalues [illegible] absolute [illegible] eigenvalues [illegible]

CAPITOLUL 9

ANALIZA DE CORELAȚIE ŞI REGRESIE

- **Introducere în analiza de corelație şi regresie**
- **Analiza de corelație**
- **Analiza de regresie**
- **Regresia multiplă în SPSS**

9.1 *Introducere în analiza de corelație și regresie*

În acest capitol vom trata problemele metodologice de bază ale studiului legăturilor statistice cu ajutorul analizei de corelație și regresie, precum și demersul specific acestor metode în SPSS.

9.1.1 Noțiunea de legătură statistică

O *legătură statistică* (stochastică) are loc atunci când modificarea unei variabile este rezultatul conjugat al influenței mai multor variabile, influență manifestată în medie, pe ansamblul unităților unei colectivități. În cazul a două variabile, X și Y, o legătură statistică are loc atunci când pentru fiecare valoare a variabilei X, variabila aleatorie Y ia valori distribuite în jurul mediei sale. Abaterile variabilei Y, în plus și minus față de medie, sunt datorate acțiunii altor variabile (altele decât variabila X).

De exemplu, în cazul legăturii dintre nivelul consumului și cel al veniturilor indivizilor unei populații, nivelul consumului depinde de nivelul veniturilor indivizilor, dar asupra consumului acționează și alți factori ale căror influențe le însumăm într-o variabilă aleatorie reziduală.

9.1.2 Probleme ale analizei de corelație și regresie

Într-o cercetare bazată pe analiza de corelație și regresie trebuie rezolvate următoarele probleme:

1. *Identificarea existenței legăturii.* Se rezolvă prin analiza logică a posibilității de existență a unei legături între variabilele considerate.
2. *Determinarea gradului de intensitate a legăturii.* Se rezolvă cu ajutorul indicatorilor parametrici sau neparametrici ai intensității corelației, folosiți în analiza de corelație.
3. *Stabilirea sensului și formei legăturii.* Se utilizează metode specifice analizei de regresie: metode elementare (serii paralele interdependente, gruparea statistică, tabelul de corelație, diagrama de tip *scatterplot*) și metode analitice (de exemplu, metoda celor mai mici pătrate).

9.2 Analiza de corelație

Analiza de corelație este folosită pentru a studia intensitatea legăturii dintre variabile. În sens strict, corelația este o măsură a intensității legăturii dintre variabile.

Legăturile statistice, în funcție de tipul variabilelor considerate, pot exprima fie asocieri (cazul variabilelor nominale), fie corelații (cazul variabilelor numerice). Ne vom opri asupra măsurării corelației.

Corelația poate fi exprimată prin: covarianță, coeficientul de corelație Pearson, raportul de corelație Pearson, coeficienți neparametrici de corelație.

9.2.1 Coeficientul de corelație Pearson

Coeficientul de corelație teoretic se notează cu $\rho(X,Y)$ și este definit de relația:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{N \cdot \sigma_x \cdot \sigma_y}, \ i = \overline{1,N},$$

unde:

– $cov(X,Y)$ – covarianța: $cov(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N}$;

– x_i, y_i *și* μ_x, μ_y – valori ale variabilelor corelate și nivelul mediu al acestora;

– N – numărul perechilor de valori;

– σ_x *și* σ_y – abaterea medie pătratică pentru *X*, respectiv *Y*.

Coeficientul de corelație este obținut prin standardizarea covarianței. Valoarea coeficientului de corelație este cuprinsă între *–1* și *+1*:

$$-1 \leq \rho \leq +1$$

Dacă ρ ia valoarea zero, atunci între variabile nu există legătură.

Semnul valorii ρ arată sensul relației dintre variabile. Semnul plus arată o legătură directă (pe măsură ce cresc valorile variabilei X, cresc și valorile variabilei Y), iar semnul minus – o legătură inversă (pe măsură ce cresc valorile variabilei X, valorile variabilei Y descresc).

Valoarea absolută a lui ρ indică intensitatea legăturii, şi anume: cu cât se apropie mai mult de 1, cu atât legătura este mai puternică, respectiv cu cât se apropie mai mult de zero, cu atât legătura este mai slabă.

Un coeficient de corelaţie egal cu +1 indică o legătură directă perfectă între variabile. Un coeficient de corelaţie egal cu – 1 arată o legătură inversă perfectă.

9.2.2 Estimarea şi testarea coeficientului de corelaţie

Un *estimator* $\hat{\rho}$ pentru coeficientul de corelaţie (ρ) are ca valori posibile *coeficienţii de corelaţie empirici* (r_{y_x}), determinaţi la nivelul eşantioanelor posibil de extras printr-o metodă de sondaj.

În acest sens, la nivelul unui eşantion de volum *n*, se determină *coeficientul de corelaţie empiric* propus de K. Pearson:

$$r_{y_x} = \frac{\text{cov}(x,y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n \cdot s_x \cdot s_y},$$

care reprezintă o estimaţie pentru parametrul ρ.

Dezvoltând relaţia de mai sus, se obţine o formulă de calcul simplificat al *coeficientului de corelaţie empiric*:

$$r_{y_x} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}, \quad i = \overline{1,n}.$$

Considerând datele cu privire la legătura dintre cantitatea de îngrăşăminte şi producţia medie de grâu la hectar, prezentate în tabelul 9.1, precum şi elementele de calcul date în tabelul 9.2 , obţinem:

$$r_{y_x} = \frac{5.420 - 15.115}{\sqrt{[5.55 - (15)^2][5.3225 - (115)^2]}} = \frac{375}{380,79} = 0,98479 .$$

Valoarea obţinută este foarte apropiată de *+1*, deci între cele două variabile există o legătură directă foarte strânsă.

Testarea semnificaţiei valorii coeficientului de corelaţie pleacă de la ipoteza că nu există corelaţie între variabile:

Ipoteza nulă $H_0 : \rho = 0$;

Ipoteza alternativă $H_1 : \rho \neq 0$.

Verificarea ipotezei H_0 se face cu ajutorul *testului t* pentru *coeficientul de corelaţie simplă.*

Testul t (*Student*) folosit pentru verificarea semnificaţiei coeficientului de corelaţie simplă este:

$$t = \frac{\hat{\rho}}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}.$$

unde:

t este o statistică *Student* cu $(n-2)$ grade de libertate;

$\hat{\sigma}_{\hat{\rho}}$ este estimatorul abaterii medii pătratice a lui $\hat{\rho}$ (estimatorul lui ρ):

$$\hat{\sigma}_{\hat{\rho}} = \sqrt{\frac{1-\hat{\rho}^2}{n-2}}.$$

La nivelul unui eşantion observat, se obţin relaţiile:

$$t_{calc} = \frac{r}{s_{\hat{\rho}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad s_{\hat{\rho}} = \sqrt{\frac{1-{r_{y_x}}^2}{n-2}}.$$

unde:

r_{y_x} – coeficientul de corelaţie simplă;

n – numărul perechilor de valori x şi y.

Valoarea calculată a lui t se compară cu valoarea teoretică obţinută din tabelul t (*Student*), pentru $n-2$ grade de libertate şi nivelul de semnificaţie stabilit.

Dacă $t_{calc} > t_{tab}$, atunci se respinge H_0 şi se trage concluzia, cu un risc considerat (de regulă, 5 %), că valoarea coeficientului de corelaţie nu este egală cu zero; respectiv, că între variabilele cercetate există o legătură semnificativă, deci coeficientul de corelaţie este semnificativ statistic.

Considerând legătura dintre cantitatea de îngrăşăminte şi producţia medie de grâu la hectar, prezentată prin datele din tabelul 9.1, cu un număr de *5* valori x şi y, pentru care s-a găsit coeficientul de corelaţie $r_{yx} = 0{,}98$, se calculează *testul t* astfel:

$$t = \frac{0{,}98\sqrt{5-2}}{\sqrt{1-0{,}98^2}} = 8{,}53.$$

În tabelul *t Student*, la $n - 2 = 3$ grade de libertate şi pentru un nivel de semnificaţie $\alpha = 0{,}01$, găsim $t = 5{,}841$.

Comparând t_{calc} cu t_{tab}, se observă că: $t_{calc} = 8{,}53 > t_{tab} = 5{,}841$, prin urmare se respinge ipoteza nulă şi se poate trage concluzia că valoarea coeficientului de corelaţie este semnificativă statistic.

9.2.3 Estimarea şi testarea raportului de corelaţie

Raportul de corelaţie Pearson este un indicator al intensităţii legăturii ce poate fi aplicat atât în cazul regresiei liniare, cât şi al celei neliniare, simple sau multiple.

Raportul de corelaţie este notat cu η şi se defineşte prin relaţiile:

$$\eta = \sqrt{\frac{\sigma^2_{y_x}}{\sigma^2_y}},$$

$$\eta = \sqrt{1 - \frac{\sigma^2_{y/y_x}}{\sigma^2_y}},$$

unde:

$\sigma^2_y = \dfrac{\sum_i (y_i - \mu_y)^2}{N}$ reprezintă varianţa generală, respectiv varianţa variabilei Y în raport cu media tuturor valorilor sale;

$\sigma^2_{y_x} = \dfrac{\sum_i (y_{x_i} - \mu_y)^2}{N}$, varianţa valorilor teoretice faţă de media lor (varianţa sub influenţa factorilor esenţiali);

$\sigma^2_{y/y_x} = \dfrac{\sum_i (y_i - y_{x_i})^2}{N}$, varianţa valorilor reale faţă de valorile teoretice ale variabilei (varianţa reziduală, eroare e_i).

Varianţa generală este egală cu *suma celorlalte două varianţe componente*:

$$\sigma^2_y = \sigma^2_{y_x} + \sigma^2_{y/y_x};$$

respectiv, variația totală (V_T) este suma variației explicate (V_E) și a variației reziduale (V_R), adică:

$$V_T = V_E + V_R.$$

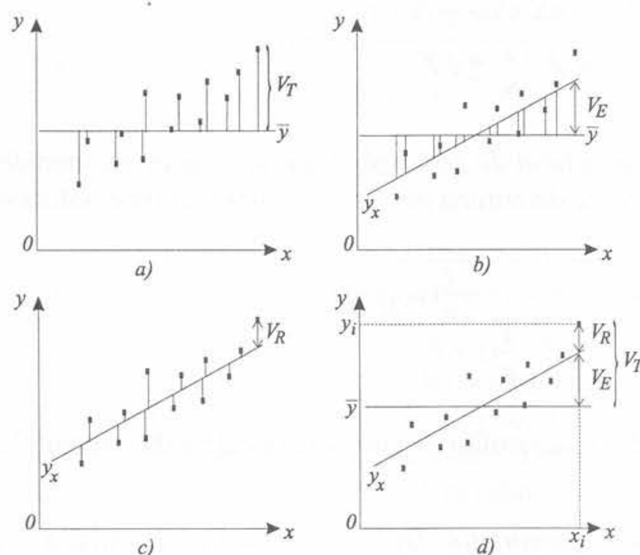


Figura 9.1 *Descompunerea pe componente a varianței unei variabile Y, într-un model de regresie: a) varianța totală; b) varianța explicată; c) varianța reziduală; d) relația între componente*

Reprezentarea grafică a descompunerii varianței unei variabile Y pe componente (varianța explicată și varianța reziduală) într-un model de regresie, precum și relația între componente sunt prezentate în figura 9.1

Valoarea raportului de corelație este un număr cuprins în intervalul: $0 \le \eta \le 1$.

Valoarea la pătrat a raportului de corelație reprezintă *raportul de determinație*:

$$\eta^2 = \frac{\sigma^2_{y_x}}{\sigma^2_y},$$

și arată ponderea influenței variabilei X asupra variației variabilei Y. Acest indicator se exprimă, de regulă, în procente, pentru a facilita interpretarea rezultatelor.

Prin explicitarea celor două varianțe, $\sigma^2_{yy_x}$ și σ^2_y, din raportul de corelație și efectuarea unor transformări elementare, se ajunge la raportul de corelație pe baza valorilor parametrilor ecuației de regresie din modelul admis.

În cazul regresiei liniare ($y = \alpha + \beta x + e$), raportul de corelație devine:

$$\eta = \sqrt{\frac{\alpha \Sigma y_i + \beta \Sigma x_i y_i - \frac{1}{N}(\Sigma y_i)^2}{\Sigma y_i^2 - \frac{1}{N}(\Sigma y_i)^2}}, \quad i = \overline{1,N}.$$

Estimarea raportului de corelație. La nivelul unui eșantion observat, raportul de corelație se poate determina pe baza valorilor empirice, folosind relația:

$$\eta_{y_x} = \sqrt{\frac{a \Sigma y_i + b \Sigma x_i y_i - \frac{1}{n}(\Sigma y_i)^2}{\Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2}}, \quad i = \overline{1,n}.$$

Aplicând relația raportului de corelație la datele din tabelul 9.1, se obține:

$\eta_{y_x} = 0{,}98$.

Raportul de determinație $(\eta^2 = 96\%)$ arată că variația variabilei *Y* este determinată în proporție de *96%* de variabila *X*; diferența până la *100%* s-ar datora factorilor aleatorii.

Dacă valoarea la pătrat a raportului de corelație $(\eta^2_{y_x})$ este egală cu valoarea la pătrat a coeficientului de corelație empiric ($r^2_{y_x}$), conform *testului B* (*Blackman*): $r^2_{y_x} = \eta^2_{y_x}$, legătura este liniară.

Testarea raportului de corelație. Testarea raportului de corelație se face pentru a verifica semnificația valorii acestuia. În acest scop este folosit *testul F* definit de relația:

$$F = \frac{n-k}{k-1} \cdot \frac{\eta^2}{1-\eta^2},$$

unde:

n – numărul valorilor observate;

k – numărul parametrilor estimați ai modelului de regresie.

Statistica F urmează o lege de distribuție *Snedecor-Fisher* de $v_1 = k - 1$ și $v_2 = n - k$ grade de libertate.

Valoarea calculată a testului se află pe baza datelor obținute la nivelul unui eșantion observat.

Dacă $F_{calc} > F_{tab}$, cu $v_1 = k - 1$ și $v_2 = n - k$ grade de libertate, atunci se trage concluzia că variabila factorială influențează semnificativ comportarea variabilei rezultative, deci raportul de corelație este semnificativ statistic.

Observație! Testul *t Student* și testul *F Fisher* conduc la rezultate identice în cazul unei regresii liniare simple.

9.2.4 Coeficienții de corelație a rangurilor

Rangul este o anumită treaptă de ordine a valorilor variabilei într-o serie. Pentru stabilirea rangurilor, valorile empirice ale variabilelor corelate sunt grupate după mărimea lor, în ordine crescătoare sau descrescătoare. De obicei, în funcție de variabila independentă se ordonează și variabila dependentă.

Coeficientul Spearman. Este o extensie a coeficientului de corelație Pearson în care valorile empirice ale variabilelor corelate sunt înlocuite cu rangurile lor corespunzătoare.

Coeficientul Spearman se notează cu θ și se calculează după relația:

$$\theta = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)},$$

unde:

d_i – reprezintă diferența dintre rangurile valorilor variabilelor corelate,

$d_i = R_{x_i} - R_{y_i}$, $i = \overline{1,n}$,

n – numărul unităților observate (numărul perechilor de valori $[y, x]$).

Coeficientul Kendall. Acest coeficient se definește prin relația:

$$\tau = \frac{S}{0,5n(n-1)} = \frac{2S}{n(n-1)},$$

unde:

$S = Q + P$, în care P reprezintă numărul de ranguri mai mari, luate în continuare față de rangul considerat, iar Q este numărul de ranguri mai mici, luate

în continuare față de rangul considerat (se ia cu semnul minus). S se calculează pentru rangurile variabilei dependente (Y), ordonate după rangurile variabilei factoriale (X);

n – numărul unităților observate.

Coeficienții de corelație a rangurilor au ca interval de variație $[-1,+1]$, cu aceeași semnificație ca și în cazul coeficientului de corelație Pearson.

9.2.5 Analiza de corelație folosind SPSS

În vederea efectuării unei analize de corelație și regresie cu ajutorul SPSS, se introduc datele în foaia *Data View*, din fișierul *Data Editor*, fiecare variabilă într-o coloană diferită. Pentru exemplificare, considerăm un caz simplu cu privire la analiza legăturii dintre cantitatea de producție la hectar și cantitatea de îngrășăminte la hectar, pe baza rezultatelor înregistrate pe un eșantion de cinci firme. Datele pentru cele două variabile sunt prezentate în tabelul 9.1.

SPSS prezintă două tipuri de corelație: *bivariată* și *parțială*. *Corelația bivariată* vizează legătura dintre două variabile, dintre care una este efectul (rezultativa, dependenta), iar cealaltă este cauza (factoriala, independenta). *Corelația parțială* prezintă corelația dintre două variabile, dintre care una este efectul controlat al influenței uneia sau a mai multor variabile factoriale.

Pentru corelația bivariată, în SPSS se pot calcula *trei coeficienți de corelație*, și anume: *Pearson*, *Kendall* și *Spearman*, precum și nivelurile de semnificație corespunzătoare unui test bilateral sau unui test unilateral.

Tabelul 9.1 *Cantitatea de îngrășăminte și producția de grâu la ha*

	firma	ingras	prod
1	a	1,00	10,00
2	b	2,00	15,00
3	c	3,00	20,00
4	d	4,00	30,00
5	e	5,00	40,00

Pentru datele considerate anterior, folosim în SPSS următorul demers: meniul *Analyze* → comanda *Correlate* → opțiunea *Bivariate*, prin care se deschide fereastra *Bivariate Correlations* (vezi figura 9.2).

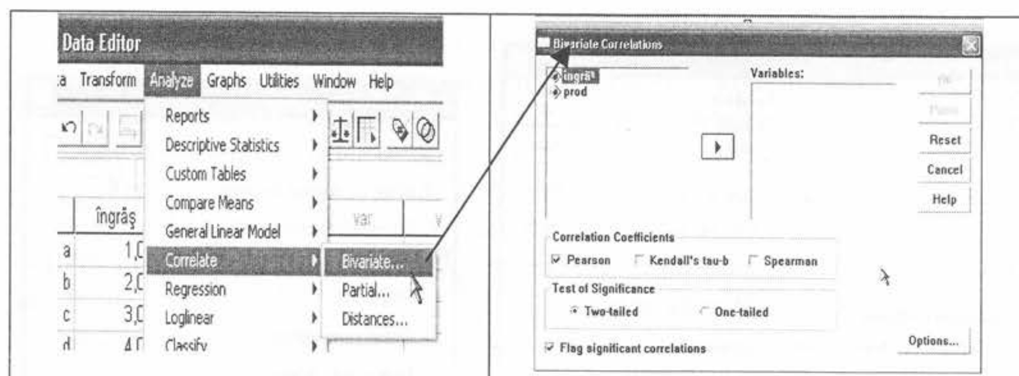


Figura 9.2 *Selectarea opțiunii* Correlate *și fereastra* Bivariate Correlations

După deschiderea ferestrei *Bivariate Correlations* se parcurg următorii pași (vezi figura 9.3):

- Selectăm variabilele dorite și le mutăm în zona *Variables*;
- În zona *Correlation Coefficients*, alegem, prin bifare în casetele de validare corespunzătoare, coeficienții de corelație pe care dorim să-i calculăm;
- În zona *Test of Significance*, alegem una din cele două opțiuni, *Two--tailed* sau *One-tailed*, care permit selectarea pragului de semnificație corespunzător ipotezelor formulate. La deschiderea ferestrei de dialog, este selectată opțiunea *Two-tailed*. Opțiunea *One-tailed* se alege atunci când se cunoaște direcția legăturii dintre cele două variabile;
- Caseta de validare *Flag significant correlations* este activată la deschiderea ferestrei dialog și are ca efect semnalizarea corelațiilor semnificative. Astfel, coeficienții de corelație semnificativi la pragul de 0,05 sunt marcați cu un asterisc, iar cei semnificativi la pragul de 0,01 sunt marcați cu două asteriscuri;
- Prin clic pe butonul de comandă *Options* deschidem fereastra *Bivariate Correlations: Options*, unde alegem opțiunile din zonele *Statistics* și *Missing Values*;
- Prin butonul de comandă *Continue* se revine în fereastra *Bivariate Correlations* din care, activând *OK*, cerem obținerea output-ului (vezi figura 9.4).

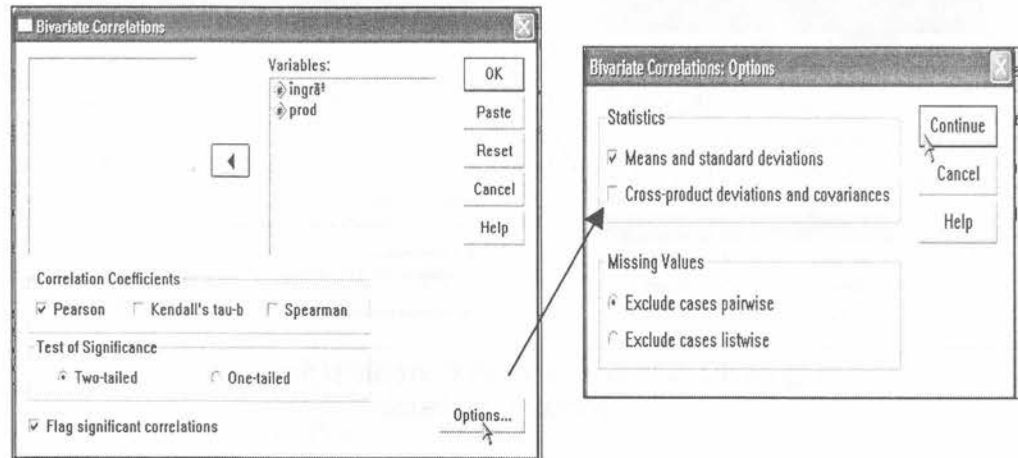


Figura 9.3 *Alegerea opțiunilor în procedeul* Corelație

În output sunt prezentate statisticile pentru fiecare variabilă, precum și valoarea coeficientului de corelație Pearson, cu nivelul de semnificație (*Sig.*) corespunzător.

Tabelul *Correlations* este un tabel cu matricea coeficienților de corelație. Valorile sunt distribuite simetric, de o parte și de alta a diagonalei coeficienților de corelație egali cu 1, corespunzători corelației fiecărei variabile cu ea însăși. De o parte și de alta a diagonalei tabelului sunt prezentate valorile coeficienților de corelație dintre variabile, luate două câte două, și valorile pragului de semnificație (*Sig.*) corespunzător, precum și numărul observațiilor considerate, *N*.

Correlations

		ÎNGRĂª	PROD
ÎNGRĂª	Pearson Correlation	1,000	,985**
	Sig. (2-tailed)	,	,002
	N	5	5
PROD	Pearson Correlation	,985**	1,000
	Sig. (2-tailed)	,002	,
	N	5	5

**. Correlation is significant at the 0.01 level

Figura 9.4 *Output SPSS pentru procedeul* Corelație

Pentru exemplul considerat s-a obţinut un *coeficient de corelaţie Pearson* egal cu *0.985*, ceea ce sugerează că între variabile există o corelaţie directă, puternică, valoarea coeficientului fiind foarte apropiată de unu (valoare corespunzătoare unei corelaţii perfecte).

Testarea semnificaţiei coeficientului de corelaţie este realizată cu ajutorul *testului t*. Valoarea *Sig.* corespunzătoare, egală cu *0,002*, evidenţiază că s-a obţinut un coeficient de corelaţie semnificativ la un prag de 0,002, adică sunt şanse mai mici de 1% de a greşi dacă afirmăm că între cele două variabile există o corelaţie semnificativă.

Observaţie! În exemplul dat, este prezentată o matrice a coeficienţilor de corelaţie bazată pe două variabile. Când se folosesc mai mult de două variabile, matricea generată de SPSS este asemănătoare, incluzând toate corelaţiile perechi posibile.

9.3 Analiza de regresie

9.3.1 Concepte şi noţiuni

Conceptul de *regresie* exprimă o legătură de tip statistic, şi anume *regresia în medie*[1] cu privire la comportamentul unor variabile.

Analiza de regresie este folosită pentru:

- estimarea valorilor unei variabile considerând valorile altei/altor variabile;
- evaluarea măsurii în care variabila dependentă poate fi explicată prin variabila independentă sau printr-un set de variabile independente;
- identificarea unui subset din mai multe variabile independente care trebuie luate în calcul pentru estimarea variabilei dependente.

Un *model de regresie*, în expresie generală, poate fi scris astfel:

$$Y = f(X_1, X_2, \ldots, X_n) + \varepsilon,$$

1. În literatura de specialitate, expresia *regresie în medie* sau *cum a scăpat omenirea de „gigantism" şi „piticism"* este legată de cercetările lui Fr. Galton asupra eredităţii. Galton, observând modul în care evoluează înălţimea copiilor faţă de cea a părinţilor, a ajuns la concluzia că, de regulă, din părinţi de talie mare se nasc copii cu o talie inferioară lor, iar din părinţi de talie redusă se nasc copii cu talie mai mare decât a părinţilor. Dacă din părinţi de talie mare s-ar fi născut copii cu talie şi mai mare ş.a.m.d., s-ar fi ajuns la gigantism sau, în celălalt caz, la piticism. Se înţelege că legitatea descoperită, ca orice legitate statistică, se verifică nu pe cazuri izolate, ci la nivelul colectivităţilor de volum mare (Fr. Galton, *Natural Inheritance*, Macmillan, London, 1889).

în care:

Y – variabila dependentă (rezultativă), aleatorie;

$X_1,..., X_n$ – variabile independente (factoriale), nonaleatorii;

ε – variabila aleatorie eroare sau reziduu.

Observație! Într-un model de regresie statistic, se adaugă termenul de eroare (ε) la ecuația de regresie, deoarece nu toate punctele de coordonate (x_i, y_i) se găsesc chiar pe linia de variație medie.

Variabila aleatorie ε însumează influențele variabilelor neincluse în model asupra variabilei Y. Variabilele aleatorii însumate în variabila ε sunt variabile *normale de medie zero*, de *aceeași varianță* σ_ε^2 (adică, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, unde $\sigma_\varepsilon^2 = M(\varepsilon_j^2)$) și sunt *independente* unele de altele.

Variabilele X și *Y* respectă *condiția de normalitate*, adică:

$$X \sim N(\mu_x, \sigma_x^2) \text{ și } Y \sim N(\mu_y, \sigma_y^2);$$

ceea ce implică:

- liniaritatea regresiei;
- normalitatea abaterilor în raport cu dreptele de regresie;
- nulitatea mediilor acestor abateri:

$$\sum_{i=1}^{n} \varepsilon_i = 0 \Rightarrow \mu_\varepsilon = (\sum_{i=1}^{n} \varepsilon_i)/n;$$

- egalitatea varianțelor lor .

Verificarea condiției de normalitate a unei distribuții este necesar a se efectua înainte de a trece la realizarea efectivă a analizei de corelație și regresie, pentru a fundamenta alegerea procedeului de tratare a legăturii dintre fenomenele considerate.

În SPSS, verificarea normalității se poate realiza prin analiza indicatorilor (media, mediana, modul, coeficientul de asimetrie – Skewness – și boltire – Kurtosis), prin analiza grafică (folosind histograma comparată cu linia curbei distribuției normale, graficul Q-Q, graficul P-P), precum și prin teste formale (testul Kolmogorov-Smirnov) (vezi capitolul 6).

Un *model de regresie simplu liniar* se poate scrie:

$$Y = \alpha + \beta X + \varepsilon.$$

Un *model de regresie multiplu liniar*, cu două sau mai multe variabile factoriale, poate fi scris:

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon .$$

Pentru valori specificate (x_i, y_i şi ε_i), se poate scrie:

$$y_i = \alpha + \beta x_i + \varepsilon_i .$$

Variabila Y condiționată de X este de medie μ_y, respectiv:

$\mu_y = \alpha + \beta X$, pentru $X = x_i$, adică:

$$y_i = M[Y / X = x_i] + \varepsilon_i = \mu_{y/x} + \varepsilon = \alpha + \beta x_i + \varepsilon_i .$$

Parametrii ecuației de regresie, într-un model de regresie simplu liniar, $Y = \alpha + \beta X + \varepsilon$, sunt:

α – ordonata la origine (valoarea variabilei Y când X = 0);

β – panta dreptei, numită şi coeficient de regresie.

Valoarea parametrului de regresie β arată gradul de dependență dintre variabile, respectiv cu cât creşte sau scade în medie variabila Y, la o modificare cu o unitate a variabilei X.

Semnul parametrului de regresie β indică direcția legăturii dintre cele două variabile corelate, şi anume:

$\beta > 0$ – legătură directă (pozitivă);

$\beta = 0$ – nu există legătură;

$\beta < 0$ – legătură inversă (negativă).

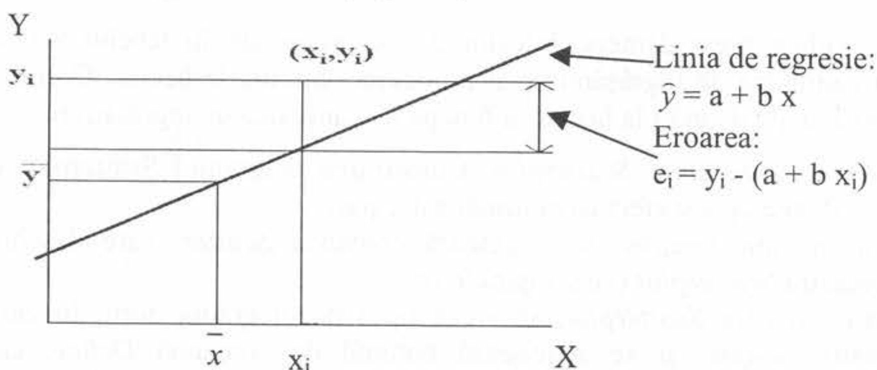


Figura 9.5 *Linia de regresie şi eroarea*

În ecuația de regresie, parametrii α și β sunt necunoscuți. În practică, parametrii unui model de regresie sunt estimați pe baza datelor unui eșantion observat, folosind estimatorul $\hat{y} = a + b x$,

unde: a și b sunt estimații ale parametrilor α și β.

Valoarea ε_i a variabilei aleatorii ε este estimată prin $e_i = y_i - (a + bx_i)$ și reprezintă distanța oricărui punct (x_i, y_i) față de linia de regresie, $\hat{y} = a + b x_i$ (vezi figura 9.5).

9.3.2 Demersul analizei de regresie

Stabilirea și analiza modelului de regresie presupun parcurgerea următorului demers:

- *construirea corelogramei*, respectiv a *norului de puncte* (*diagrama de dispersie* sau *Scatterplot*);
- *aproximarea formei legăturii* printr-un model de regresie și scrierea ecuației corespunzătoare. Se pot folosi *metode tabelare* și *metode grafice*. De regulă, pentru aproximarea modelului de regresie, adică a modelului care exprimă cel mai bine relația dintre variabile, se ajustează vizual diagrama *Scatterplot*;
- *estimarea parametrilor ecuației de regresie* (pe baza metodei celor mai mici pătrate) și interpretarea regresiei în funcție de semnul și valoarea parametrilor modelului de regresie;
- *testarea semnificației parametrilor de regresie.*

9.3.3 Aproximarea modelului de regresie folosind SPSS

Pentru a explica acest demers, folosim datele prezentate în tabelul 9.1, cu privire la cantitatea de îngrășăminte și producția obținută la hectar. Dorim să estimăm producția de grâu la hectar în funcție de cantitatea de îngrășăminte.

Construirea diagramei Scatterplot. Construirea diagramei Scatterplot cu ajutorul *SPSS* presupune efectuarea următorilor pași:

- Din meniul *Graphs*, se selectează comanda *Scatter*, care deschide fereastra *Scatterplot* (vezi figura 9.6);
- Din fereastra *Scatterplot*, se alege tipul de diagramă dorit, în cazul nostru *Simple*, și se acționează butonul de comandă *Define*, care deschide fereastra *Simple Scatterplot* pentru a defini elementele pe baza cărora SPSS va realiza diagrama;

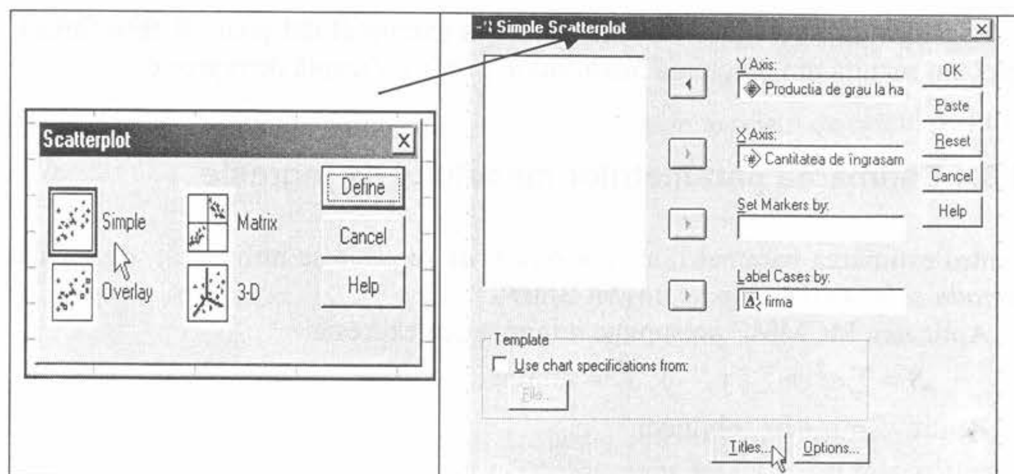


Figura 9.6 *Ferestre de dialog* Scatterplot

- În fereastra *Simple Scatterplot*, selectăm variabilele considerate și le mutăm în zonele corespunzătoare, și anume: variabila dependentă în zona *Y Axis*, variabila independentă în zona *X Axis*, iar numele unităților observate în zona *Label Cases by*;
- Cu ajutorul butoanelor de comandă *Titles* și *Options* se stabilesc liniile pentru titlu și subtitlu și, respectiv, se definesc anumite opțiuni;

Activând butonul *OK* se comandă obținerea output-ului (vezi figura 9.7).

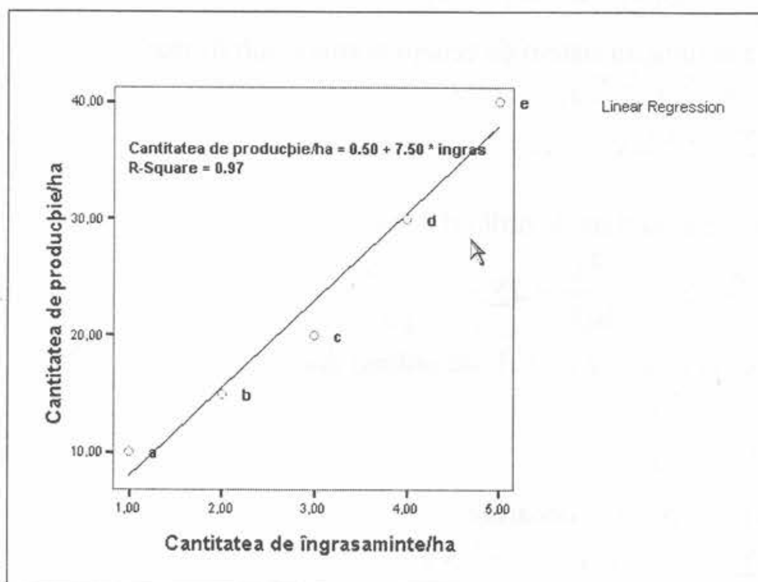


Figura 9.7 *Legătura dintre cantitatea de îngrășăminte și producția de grâu la ha*

Legătura dintre variabilele considerate în exemplul dat poate fi aproximată, aşa cum rezultă din diagrama *Scatterplot*, printr-o dreaptă de regresie.

9.3.4 Estimarea parametrilor modelului de regresie

Pentru estimarea parametrilor unui model de regresie se utilizează, de regulă, *metoda celor mai mici pătrate (MCMMP)*.

Aplicarea MCMMP presupune minimizarea expresiei:

$$S = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \text{minim}.$$

Pentru $\hat{y} = a + bx$, obţinem:

$$S = \sum (y_i - a - bx_i)^2 = \text{minim}.$$

Rezolvarea problemei de minim impune două condiţii:

1. anularea derivatelor parţiale de ordinul întâi ale lui S în raport cu a şi b;
2. matricea derivatelor parţiale de ordinul doi să fie definită pozitiv.

1. Derivatele parţiale de ordinul întâi se obţin pe baza relaţiilor:

$$\frac{\partial S}{\partial a} = 2\sum (y_i - a - bx_i)(-1) = 0,$$

$$\frac{\partial S}{\partial b} = 2\sum (y_i - a - bx_i)(-x_i) = 0, \quad \text{i} = \overline{1,\text{n}},$$

din care rezultă un sistem de ecuaţii normale sub forma:

$$na + b\sum x_i = \sum y_i$$

$$a\sum x_i + b\sum x_i^2 = \sum x_i y_i, \quad \text{i} = \overline{1,\text{n}}.$$

2. Derivatele parţiale de ordinul doi sunt:

$$\frac{\partial^2 S}{\partial a^2} = 2n, \quad \frac{\partial^2 S}{\partial a \partial b} = 2\sum x_i, \quad \frac{\partial^2 S}{\partial b^2} = 2\sum x_i^2$$

Matricea derivatelor parţiale de ordinul doi

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

este definită pozitiv, deoarece:

$$n\sum x_i^2 - (\sum x_i)^2 = n\sigma^2 > 0.$$

Rezolvarea sistemului de ecuaţii normale, printr-una din metodele cunoscute (metoda determinanţilor, metoda Doolittle etc.), conduce la obţinerea estimaţiilor a şi b ale parametrilor modelului de regresie.

Aplicând metoda determinanţilor, rezultă următoarele relaţii de calcul pentru a şi b:

$$b = \frac{\Delta b}{\Delta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \quad i = \overline{1,n};$$

$$a = \frac{\Delta a}{\Delta} = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2};$$

$$a = \bar{y} - b\bar{x}.$$

Observaţie! a şi b reprezintă valori de sondaj, *estimaţii ale parametrilor* α şi β, calculate la nivelul unui eşantion, prin aplicarea metodei celor mai mici pătrate.

Estimarea prin interval de încredere a parametrilor α *şi* β. Estimarea prin interval de încredere se bazează pe distribuţiile de selecţie ale estimatorilor $\hat{\alpha}$ şi $\hat{\beta}$ ai parametrilor α şi β.

Pentru modelul liniar simplu, se poate demonstra că *estimatorii parametrilor urmează o lege de distribuţie normală* şi sunt *nedeplasaţi*:

$$\hat{\alpha} \sim N(\alpha, \sigma_{\hat{\alpha}}^2); \quad M(\hat{\alpha}) = \alpha; \quad V(\hat{\alpha}) = \sigma_{\hat{\alpha}}^2; \quad \sigma_{\hat{\alpha}}^2 = \frac{\sum_i X_i^2}{n\sum_i (X_i - \bar{X})^2}\sigma_e^2;$$

$$\hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2); \quad M(\hat{\beta}) = \beta; \quad V(\hat{\beta}) = \sigma_{\hat{\beta}}^2; \quad \sigma_{\hat{\beta}}^2 = \frac{\sigma_e^2}{\sum_i (X_i - \bar{X})^2}.$$

Estimaţiile pentru varianţa estimatorilor parametrilor α şi β, respectiv pentru varianţa erorilor, se calculează după relaţiile:

– *varianţa estimatorului* $\hat{\alpha}$:

$$s_{\hat{\alpha}}^2 = \frac{\sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2} s_e^2;$$

– *varianța estimatorului* $\hat{\beta}$:

$$s_{\hat{\beta}}^2 = \frac{s_e^2}{\sum_i (x_i - \bar{x})^2};$$

– *varianța erorilor* (σ_e^2):

$$s_e^2 = \frac{\sum_i e_i^2}{n-2} = \frac{\sum_i (y_i - a - bx_i)^2}{n-2}.$$

Intervalul de încredere pentru coeficienții de regresie α și β, estimat pentru un eșantion observat, este definit de relațiile:

$$a \pm t_{\alpha/2} \cdot s_{\hat{\alpha}};$$

$$b \pm t_{\alpha/2} \cdot s_{\hat{\beta}}.$$

Estimațiile punctuale a și b pentru coeficienții de regresie α și β se află pe baza elementelor de calcul din tabelul 9.2.

Ecuația estimată este:

$$\hat{y} = a + bx = 0{,}5 + 7{,}5x.$$

Tabelul 9.2 *Elemente de calcul*

x_i	y_i	x_i^2	x_iy_i	y_i^2	y_{x_i}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	e_i	e_i^2
1	2	3	4	5	6	7	8	9	10
1	10	1	10	100	8,0	-2	4	2	4
2	15	4	30	225	15,5	-1	1	-0,5	0,25
3	20	9	60	400	23,0	0	0	-3,0	9,00
4	30	16	120	900	30,5	1	1	-0,5	0,25
5	40	25	200	1600	38,0	2	4	2	4,00
15	115	55	420	3225	115,0	-	10	0,0	17,50

Calculul estimației varianței erorii.

Considerând datele din tabelul 9.1, s-au calculat: $b = 7{,}5$ și $\sum(x_i - \bar{x})^2 = 10$ (coloanele 7 și 8 din tabelul 9.2). Valorile $s_{\hat{\beta}}$ și s_e se pot calcula pe baza elementelor din tabelul 9.2, coloanele 9 și 10.

Estimația varianței erorii este:

$$s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{17{,}5}{5-2} = 5{,}83.$$

Estimația varianței estimatorului $\hat{\beta}$ este:

$$s_{\hat{\beta}}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{5{,}83}{10} = 0{,}583; \quad s_{\hat{\beta}} = 0{,}76376 \quad \text{(vezi tabelul 9.6).}$$

Calculul *testului t Student*:

$$t_{calc} = \frac{b}{s_{\hat{\beta}}} = \frac{7{,}5}{0{,}76376} = 9{,}8198.$$

Pentru exemplul considerat, a rezultat o valoare $t_{calc} = 9{,}82$ (vezi tabelul 9.6), iar pentru valoarea teoretică citim din tabelul *Student*, pentru $\alpha/2 = 0{,}025$ și $n\text{-}2=3$, $t_{0,025;3} = 3{,}182$. Ca urmare, pentru $t_{calc} > t_{0,025;3}$, coeficientul de regresie β este semnificativ diferit de 0, adică variabila X explică variabila Y.

Determinarea intervalului de încredere. Intervalul de încredere pentru *coeficientul de regresie* β, considerând un risc $\alpha = 0{,}05$, este prezentat în figura 9.8 și este definit de relația:

$$b \pm t_{\alpha/2} \cdot s_{\hat{\beta}}.$$

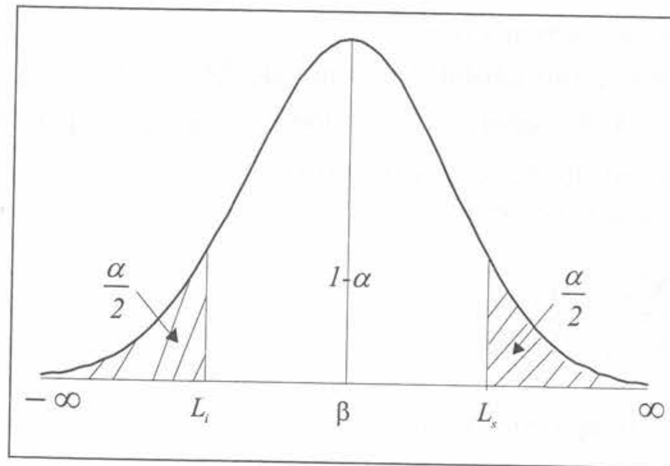


Figura 9. 8 *Distribuția de selecție a estimatorului* $\hat{\beta}$ *și intervalul de încredere*

Astfel, folosind datele din exemplul considerat anterior, pentru un risc $\alpha = 0,05$, rezultă, la nivelul eșantionului observat, următorul interval de încredere:

$$I.C. = (7,5 \pm 0,76376 \cdot 3,182) = [5,07; 9,93].$$

Putem spune că ne asumăm un risc de *5%* ca valoarea adevărată a coeficientului de regresie β să nu fie acoperită de intervalul [*5,07; 9,93*].

Dacă intervalul de încredere pentru β ar conține valoarea *0*, atunci nu s-ar respinge ipoteza H_0, ceea ce nu este cazul în exemplul nostru, deci factorul *X* influențează semnificativ variabila *Y*.

9.3.5 Estimarea parametrilor modelului de regresie folosind SPSS

Procesul de estimare a parametrilor unui model de regresie în SPSS este cunoscut ca *fitting the model* și presupune parcurgerea demersului: meniul *Analyze* → comanda *Regression* → opțiunea *Linear*, prin care se deschide fereastra de dialog *Linear Regression* (vezi figura 9.9).

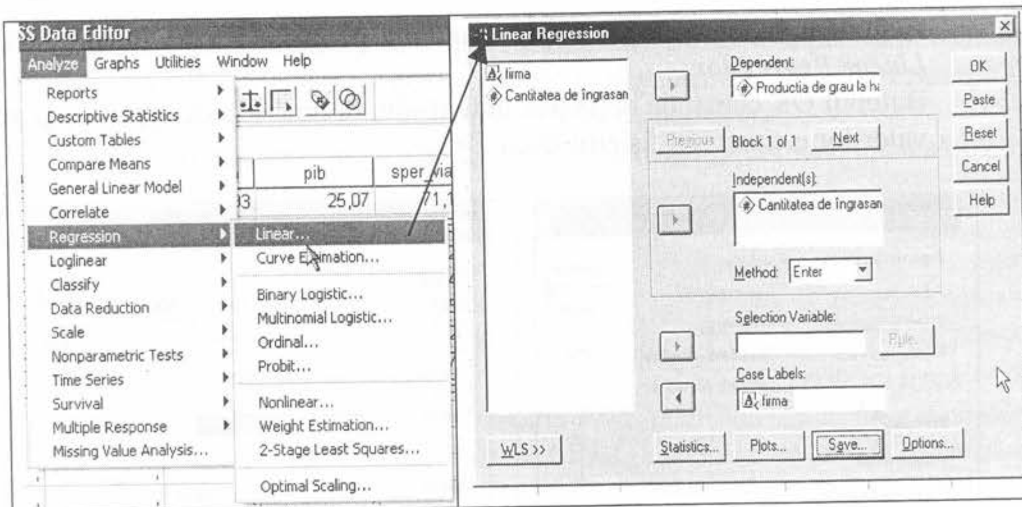


Figura 9.9 *Fereastra de dialog* Linear Regression

În fereastra dialog *Linear Regression* selectăm variabilele considerate și le mutăm în zonele de lucru corespunzătoare. În exemplul nostru (vezi tabelul 9.1), selectăm variabila rezultativă *prod* și o mutăm în zona *Dependent*, iar variabila factorială *ingras* – în zona *Independent*. În zona *Case Labels* mutăm *firma*.

În continuare se parcurg următorii pași:

- Alegem din lista *Method*, ca metodă de lucru, opțiunea *Enter*;
- Activăm butonul de comandă *Statistics* care deschide fereastra de dialog *Linear Regression: Statistics* în care bifăm casetele de validare: *Estimates*, *Confidence intervals*, *Model fit* și *Descriptives* (vezi figura 9.10);
- Butonul de comandă *Continue* determină revenirea în fereastra *Linear Regression* în care activăm butonul *Plots*, care deschide fereastra *Linear Regression: Plots*;
- În fereastra de dialog *Linear Regression: Plots*, selectăm și mutăm SRESID în zona *Y*, respectiv ZPRED în zona *X*. Pentru *Standardized Residual Plots*, bifăm casetele de validare *Histogram* și *Normal probability plot*;
- Butonul de comandă *Continue* determină revenirea în fereastra *Linear Regression* în care activăm butonul *Save*;
- În fereastra *Linear Regression: Save* (vezi figura 9.11), pentru *Predicted Values* bifăm caseta *Unstandardized*, pentru *Prediction Intervals* bifăm caseta *Mean*, iar pentru *Residuals* alegem *Unstandardized*;

- Acționăm butonul de comandă *Continue* pentru a reveni în fereastra *Linear Regression*;
- Butonul *OK* comandă obținerea output-ului în fereastra de rezultate și a valorilor estimate în fișierul *Data Editor*.

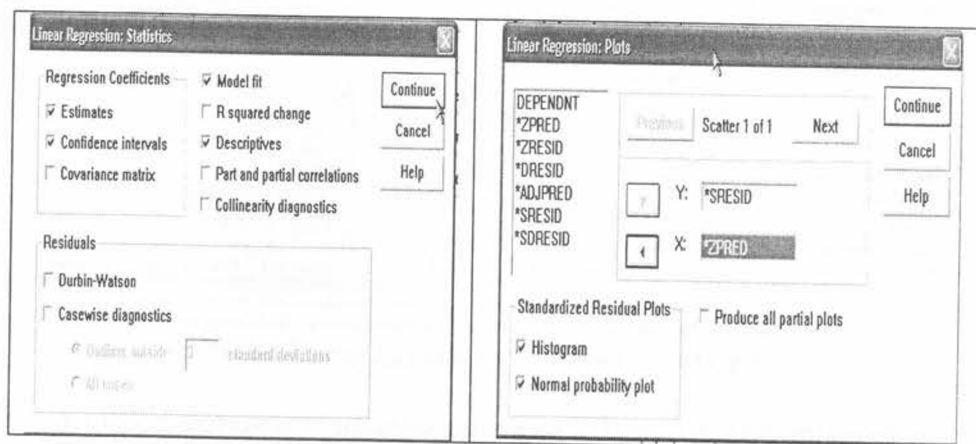


Figura 9.10 *Ferestrele* Statistics *și* Plots *pentru un model de regresie*

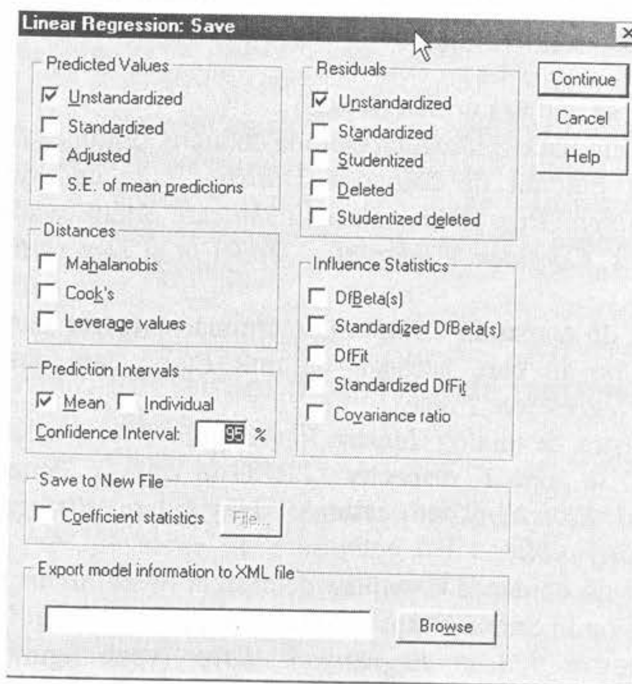


Figura 9.11 *Fereastra dialog* Linear Regression: Save

În fișierul *Data Editor*, în foaia *Data View*, SPSS completează coloane distincte cu *valorile estimate* pentru *variabila dependentă*, *valorile reziduale* și *limitele* inferioară și superioară ale *intervalului de încredere*.

Pentru exemplul considerat, rezultatele estimării sunt prezentate în tabelul 9.3.

Tabelul 9.3 *Valori estimate pentru producția de grâu la ha, pe baza eșantionului de firme prezentat în tabelul 9.1*

firma	ingras	prod	pre_1	res_1	lmci_1	umci_1
a	1.00	10.00	8.00000	2.00000	2.04619	13.95381
b	2.00	15.00	15.50000	-.50000	11.29002	19.70998
c	3.00	20.00	23.00000	-3.00000	19.56257	26.43743
d	4.00	30.00	30.50000	-.50000	26.29002	34.70998
e	5.00	40.00	38.00000	2.00000	32.04619	43.95381

Fereastra de rezultate – Output-ul – pentru analiza de regresie conține: *Model Summary*, *ANOVA*, *Coefficients*, *Normal P-P plot* și *Scatterplot*.

Tabelul *Model Summary* prezintă valoarea coeficientului de corelație (R), valoarea raportului de determinație (R^2), valoarea ajustată a lui R și eroarea standard a estimației. Pentru exemplul considerat, *Model Summary* este prezentat în tabelul 9.4 (vezi și output-ul din figura 9.4)

Tabelul 9.4 Model Summary, *cazul regresiei simple*

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.985	.970	.960	2.4152

a Predictors: (Constant), Cantitatea de ingrasaminte la ha

b Dependent Variable: Productia de grau la ha

Valoarea R arată dacă există sau nu o corelație între variabila dependentă (rezultativa Y) și variabila independentă (factoriala X). Acest indicator ia valori între –1 și 1 (vezi paragrafele 9.2.1 și 9.2.2).

Interpretarea modelului. În interpretarea modelului, se folosește coeficientul de determinație, R^2.

Raportul de determinație R^2 arată proporția variației variabilei dependente explicate prin modelul de regresie şi este folosit pentru a evalua calitatea ajustării (alegerea modelului).

R^2 ia valori între 0 şi 1. Dacă R^2 este egal cu 0 sau are o valoare foarte mică, atunci modelul de regresie ales nu explică legătura dintre variabile; relația dintre variabila dependentă şi variabila independentă nu coincide cu modelul ales, de exemplu, liniar. Dacă R^2 este egal cu 1, atunci toate observațiile cad pe linia de regresie, deci modelul de regresie explică perfect legătura dintre variabile. Ca urmare, R^2 este folosit pentru a stabili care model de regresie este cel mai bun. Această metodă de alegere a modelului de regresie potrivit este recomandată pentru modelele care nu conțin un număr mare de variabile.

Pentru exemplul considerat, a rezultat o valoare $R=0,985$, respectiv, $R^2= 0,970$, ceea ce ne arată că între cantitatea de producție/ha şi cantitatea de îngrăşăminte/ha există o legătură liniară, directă, foarte strânsă (vezi tabelul 9.4).

Tabelul *Regression ANOVA* prezintă rezultatele analizei varianței variabilei dependente sub influența factorului de regresie şi a factorului reziduu. Adică prezintă informații asupra sumei pătratelor abaterilor variabilei dependente, datorate modelului de regresie şi factorului reziduu, gradele de libertate, estimațiile varianțelor datorate celor două surse de variație (regresie şi reziduu), raportul F şi *Sig.* (vezi tabelul 9.5).

Tabelul 9.5 ANOVA *pentru regresie*

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	562.500	1	562.500	96.429	.002
	Residual	17.500	3	5.833		
	Total	580.000	4			

a Predictors: (Constant), Cantitatea de îngrasaminte la ha
b Dependent Variable: Productia de grau la ha

Statistica test F se obține ca raport între media pătratelor abaterilor datorate regresiei şi media pătratelor abaterilor datorate reziduului, calculate cu gradele de libertate corespunzătoare. Această statistică test este folosită pentru testarea modelului de regresie, adică a ipotezei prin care se presupune că panta dreptei (β_1) este 0, respectiv, pentru regresia multiplă, $\beta_1, ...\beta_p = 0$.

Dacă *testul F* ia o valoare mare, iar valoarea *Sig.* corespunzătoare statisticii *F* este mică (mai mică decât 0,05), atunci variabila independentă explică variația variabilei dependente și invers.

În exemplul considerat, valoarea *Sig.* pentru *F* este mai mică decât 0,05, deci relația liniară dintre cele două variabile considerate este semnificativă (vezi tabelul 9.5).

Tabelul 9.6 *Coeficienții de corelație*

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	.500	.533		.197	.856
	Cantitatea de ingrasaminte la ha	7.500	.764	.985	.820	.002

a Dependent Variable: Productia de grau la ha

Coeficienții de regresie. Tabelul *Coefficients* (vezi Tabelul 9.6) prezintă coeficienții nestandardizați ai modelului de regresie estimat, erorile standard ale acestora, coeficienții de regresie standardizați cu erorile standard corespunzătoare, precum și valorile statisticii test *t* și valorile *Sig.* corespunzătoare.

Coeficienții de regresie standardizați sunt folosiți atunci când într-un model intră mai multe variabile independente exprimate în unități de măsură diferite, în scopul facilitării comparării acestora.

Testarea parametrilor modelului de regresie se face cu ajutorul testului *t*, pentru a afla care este probabilitatea ca fiecare parametru să fie nul :

$$H_0 : \beta = 0$$

Pentru exemplul dat, valoarea Sig.=0.002 este mai mică decât 0.05, arătând că β (panta dreptei de regresie) corespunde unei legături semnificative între cele două variabile.

9.4 Regresia multiplă în SPSS

9.4.1 Modelul de regresie multiplă

Un model statistic de regresie multiplă este definit de relația:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon,$$

unde:

Y este variabila dependentă ;

$X_1, X_2, ..., X_p$ sunt variabile independente (predictori);

ε este variabila aleatorie eroare (reziduu);

α, β_i sunt coeficienții de regresie.

9.4.2 Selecția variabilelor independente într-un model de regresie

Pentru a găsi cea mai bună combinație de variabile independente care explică variația variabilei dependente, într-un model de regresie, SPSS oferă mai multe metode: *Forward*, *Backward*, *Stepwise*. Prin aceste metode se pot selecta variabilele care explică optim variația variabilei dependente. Aplicarea lor presupune introducerea și eliminarea variabilelor independente în model în funcție de gradul de semnificație a legăturii lor cu variabila dependentă, până când nici o variabilă nu mai poate fi introdusă sau eliminată din ecuația de regresie.

1. *Forward* (*introducerea pas cu pas*). Prin acest procedeu, variabilele independente sunt introduse în model una câte una (pas cu pas), în ordinea importanței lor. În pasul întâi, este introdusă variabila care este cel mai puternic corelată, pozitiv sau negativ, cu variabila dependentă. În pasul doi (și următorii), se introduc variabile mai puțin corelate. La fiecare pas este testată ipoteza de nul asupra coeficientului de regresie a variabilei introduse, adică se testează dacă coeficientul de regresie corespunzător este zero. Este folosită statistica *test t* (respectiv, statistica F care este pătratul statisticii t). Pașii se opresc când un prag de semnificație stabilit pentru F nu mai este atins.

2. *Backward* (*eliminarea pas cu pas*). Acest procedeu este cel mai des folosit în practică. Începe cu toate variabilele considerate în model și la fiecare pas se elimină cel mai slab predictor (variabilă independentă). Cel mai slab predictor este definit de variabila independentă cel mai puțin importantă, adică

variabila care determină cea mai mică reducere a statisticii Fisher, *F*. Variabilele sunt eliminate până când un prag de semnificaţie stabilit pentru *F* nu mai este atins.

3. *Stepwise* (*selecţia pas cu pas*). Acest procedeu începe la fel ca Forward, dar la fiecare pas testează variabilele existente deja în model, pentru a le elimina. Aceasta este metoda cea mai folosită, în special când există corelaţii între variabilele independente. De exemplu, introducerea celei de-a patra variabile poate diminua importanţa unei variabile deja introduse şi, ca urmare, aceasta este eliminată din model (în Forward aceasta rămâne în model).

9.4.3 Exemplu de regresie multiplă folosind SPSS

Pentru realizarea în SPSS a unei analize de regresie multiplă, vom considera datele din *reg_pib_inv_cs_pocup.sav*, referitoare la regiunile României în anul 2000 şi procedeul *Backward*.

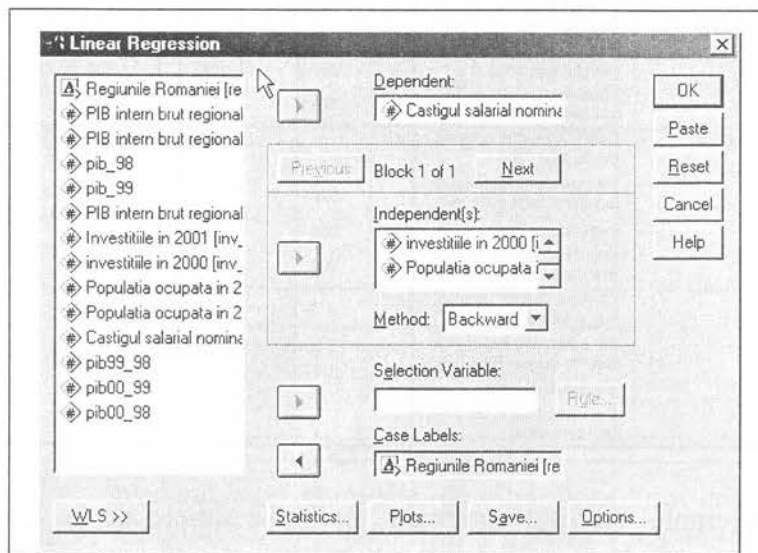


Figura 9.12 *Fereastra de dialog* Linear Regression, *metoda* Backward

Paşii demersului sunt cei prezentaţi în cazul unui model de regresie liniară simplă, cu elementele specifice unui model de regresie multiplă. Astfel, în fereastra *Linear Regression* selectăm (vezi figura 9.12):

- în zona *Dependent*: câştigul salarial nominal net (*cs*);
- în zona *Independent(s)*: produsul intern brut regional (*pib*), investiţiile în 2000 (*inv*) şi populaţia ocupată în 2000 (*pocup*);

- în zona *Method*: se alege metoda *Backward*;
- în zona *Case Labels*: regiunile României (*reg*);
- în fereastra *Linear Regression: Statistics*, deschisă prin butonul de comandă *Statistics*, se activează casetele de validare *Descriptives* şi *Collinearity diagnostics*.

În tabelul *Correlations* se afişează coeficienţii de corelaţie Pearson (*Pearson Correlation*), valoarea semnificaţiei (*Sig.*) pentru fiecare coeficient de corelaţie şi numărul cazurilor considerate în studiu (*N*).

Tabelul 9.7 *Matricea corelaţiilor parţiale*

Correlations

		Castigul salarial nominal net in anul 2000	PIB intern brut regional pe locuitor in anul 2000 (lei)	investitiile in 2000	Populatia ocupata in 2000 (mii persoane)
Pearson Correlation	Castigul salarial nominal net in anul 2000	1.000	.001	.877	-.555
	PIB intern brut regional pe locuitor in anul 2000 (lei)	.001	1.000	.157	-.710
	investitiile in 2000	.877	.157	1.000	-.737
	Populatia ocupata in 2000 (mii persoane)	-.555	-.710	-.737	1.000
Sig. (1-tailed)	Castigul salarial nominal net in anul 2000	.	.500	.005	.098
	PIB intern brut regional pe locuitor in anul 2000 (lei)	.500	.	.368	.037
	investitiile in 2000	.005	.368	.	.029
	Populatia ocupata in 2000 (mii persoane)	.098	.037	.029	.
N	Castigul salarial nominal net in anul 2000	7	7	7	7
	PIB intern brut regional pe locuitor in anul 2000 (lei)	7	7	7	7
	investitiile in 2000	7	7	7	7
	Populatia ocupata in 2000 (mii persoane)	7	7	7	7

Pentru exemplul dat sunt prezentate corelaţiile simple ale fiecărei variabile independente (predictor) cu variabila dependentă *cs* – câştigul salarial nominal net (vezi matricea corelaţiilor din Tabelul 9.7).

Se observă că valoarea coeficienţilor de corelaţie de pe diagonală este egală cu 1, deoarece fiecare variabilă este corelată perfect cu ea însăşi. Se constată că legătura cea mai semnificativă este între câştigul salarial nominal net şi investiţii. Între variabila dependentă – *cs* – şi variabila independentă – *inv* – există o legătură directă, puternică. Valoarea coeficientului de corelaţie este egală cu 0,877, cu o valoare *Sig.* mai mică decât 0,05.

Tabelul *Variable Entered/Removed* furnizează o prezentare a rezultatelor eliminării pas cu pas a variabilelor (vezi tabelul 9.8).

SPSS elaborează, la început, un model cu toate variabilele independente, folosind metoda *Enter*, apoi, în fiecare pas, creează un model, eliminând variabila care are cea mai redusă contribuţie.

Tabelul 9.8 *Variabilele introduse în model şi variabilele eliminate pas cu pas*

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Populatia ocupata in 2000 (mii persoane), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000[a]	.	Enter
2	.	Populatia ocupata in 2000 (mii persoane)	Backward (criterion: Probability of F-to-remove >= .100).
3	.	PIB intern brut regional pe locuitor in anul 2000 (lei)	Backward (criterion: Probability of F-to-remove >= .100).

a. All requested variables entered.

b. Dependent Variable: Castigul salarial nominal net in anul 2000

În exemplul considerat, sunt eliminate, pe rând, în ordinea celei mai slabe influenţe asupra câştigului salarial nominal net, variabila *populaţie ocupată* şi variabila *produs intern brut pe locuitor*.

Tabelul *Model Summary* prezintă pentru fiecare model de regresie valoarea coeficientul de corelaţie (R), valoarea coeficientului de determinaţie (R^2) şi eroarea standard. Valoarea R^2 creşte pe măsură ce se introduc mai multe variabile în model. Includerea de variabile irelevante duce, de asemenea, la creşterea erorii standard.

Tabelul 9.9 Model Summary, *cazul regresiei multiple*

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.888[a]	.789	.578	130351.7763	.789	3.745	3	3	.153
2	.888[b]	.789	.683	113053.5944	-.001	.009	1	5	.931
3	.877[c]	.769	.723	105627.4703	-.019	.365	1	6	.578

a. Predictors: (Constant), Populatia ocupata in 2000 (mii persoane), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000

b. Predictors: (Constant), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000

c. Predictors: (Constant), investitiile in 2000

d. Dependent Variable: Castigul salarial nominal net in anul 2000

În exemplul dat, valoarea R, valoarea R^2 ajustat și eroarea standard arată că cel mai bun predictor (variabila independentă care estimează cel mai bine variabila dependentă) este variabila „investiții".

Aceeași concluzie se poate trage considerând rezultatele din tabelul ANOVA (vezi tabelul 9.10). Dacă valoarea semnificației statisticii F este mică (*Sig.* este mai mică decât 0,05), atunci variabilele independente explică variația variabilei dependente. Cea mai mică valoare *Sig.* corespunde modelului care explică variația câștigului salarial nominal net în funcție de investiții.

În tabelul *coeficienților de regresie*, în prima parte apar coeficienții de regresie, erorile standard, valoarea statisticii *test t* pentru fiecare coeficient, precum și valoarea *Sig.* În cazul unei regresii multiple, apar, în plus față de cazul unei corelații simple, statisticile de coliniaritate (*collinearity statistics*), toleranța (*tolerance*) și factorul de inflație a varianței (*variance inflation factor* – VIF).

Coliniaritatea exprimă existența unei corelații puternice între variabilele independente. În astfel de situații se calculează statisticile toleranței; considerând numai variabilele independente, variabila dependentă este exclusă din model.

Tabelul 9.10 ANOVA

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.91E+11	3	6.363E+10	3.745	.153[a]
	Residual	5.10E+10	3	1.699E+10		
	Total	2.42E+11	6			
2	Regression	1.91E+11	2	9.536E+10	7.461	.045[b]
	Residual	5.11E+10	4	1.278E+10		
	Total	2.42E+11	6			
3	Regression	1.86E+11	1	1.861E+11	16.677	.010[c]
	Residual	5.58E+10	5	1.116E+10		
	Total	2.42E+11	6			

a. Predictors: (Constant), Populatia ocupata in 2000 (mii persoane), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000

b. Predictors: (Constant), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000

c. Predictors: (Constant), investitiile in 2000

d. Dependent Variable: Castigul salarial nominal net in anul 2000

Toleranța fiecărei variabile X_i se calcuează după relația:

Toleranța = $1 - R_i^2$,

unde:

R_i^2 este pătratul coeficientului de corelație multiplă a variabilei X_i cu toate celelalte variabile independente.

VIF este reciproca toleranței.

Toleranța poate lua valori de la 0 la 1. Cu cât valoarea toleranței este mai mică, mai apropiată de zero, cu atât variabila independentă X_i este explicată printr-o combinație liniară a celorlalte variabile independente. Ca urmare, explicarea variabilei dependente prin această variabilă poate fi considerată ca având prea puțină acuratețe.

Tabelul 9.11 *Coeficienții de regresie*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	2191707	2170142		1.010	.387	-4714653.203	9098066.708		
	PIB intern brut regional pe locuitor in anul 2000 (lei)	-4.01E-03	.026	-.091	-.155	.887	-.087	.079	.202	4.954
	investitiile in 2000	42.296	27.310	.951	1.549	.219	-44.615	129.208	.186	5.367
	Populatia ocupata in 2000 (mii persoane)	81.670	870.088	.081	.094	.931	-2687.337	2850.677	.095	10.566
2	(Constant)	2391840	350602.0		6.822	.002	1418412.850	3365267.294		
	PIB intern brut regional pe locuitor in anul 2000 (lei)	-6.18E-03	.010	-.141	-.604	.578	-.035	.022	.975	1.025
	investitiile in 2000	39.991	10.352	.899	3.863	.018	11.248	68.733	.975	1.025
3	(Constant)	2210270	168505.7		13.117	.000	1777112.723	2643428.213		
	investitiile in 2000	39.010	9.552	.877	4.084	.010	14.454	63.565	1.000	1.000

a. Dependent Variable: Castigul salarial nominal net in anul 2000

Diagnosticul coliniarității presupune analiza rezultatelor din tabelul *Collinearity Diagnostics* (vezi tabelul 9.12).

Tabelul 9.12 *Diagnosticul coliniarității*

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	PIB intern brut regional pe locuitor in anul 2000 (lei)	investitiile in 2000	Populatia ocupata in 2000 (mii persoane)
1	1	3.906	1.000	.00	.00	.00	.00
	2	7.156E-02	7.388	.00	.00	.08	.02
	3	2.237E-02	13.214	.00	.10	.10	.02
	4	3.752E-04	102.030	1.00	.90	.83	.97
2	1	2.954	1.000	.00	.00	.01	
	2	3.752E-02	8.873	.05	.09	.97	
	3	8.423E-03	18.727	.95	.91	.02	
3	1	1.972	1.000	.01		.01	
	2	2.847E-02	8.321	.99		.99	

a. Dependent Variable: Castigul salarial nominal net in anul 2000

Eigenvalue dă o indicație asupra numărului de legături care există între variabilele independente. Când mai multe *eigenvalues* sunt apropiate de zero, variabilele sunt puternic intercorelate.

Indicii de condiție se calculează ca rădăcină pătrată din raportul dintre valoarea *eigenvalue* cea mai mare și valoarea *eigenvalue* a fiecărei dimensiuni. Un indice mai mare de 15 arată că există o posibilă problemă de coliniaritate, iar o valoare mai mare de 30 indică probleme grave de coliniaritate. Aceste situații le întâlnim în exemplul considerat: pentru modelul 1, indicele corespunzător dimensiunii 4 (variabila „populația ocupată") are valoarea de 102,030, respectiv, pentru modelul 2, indicele corespunzător dimensiunii 3 (variabila „PIB") are valoarea de 18,721 (vezi tabelul 9.12).

Proporția varianței evidențiază contribuția fiecărei variabile la varianță. Variabilele care au valori mari pentru acest indicator arată probleme de coliniaritate. În exemplul dat, variabilele cu probleme de coliniaritate și care influențează substanțial varianța sunt:

- populația ocupată, cu o proporție de 0,97;
- PIB regional, cu o proporție de 0,91.

Tabelul *Excluded Variables* prezintă informații despre variabilele care sunt excluse la fiecare pas (vezi tabelul 9.13).

Tabelul 9.13 *Variabile excluse*

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
2	Populatia ocupata in 2000 (mii persoane)	.081[a]	.094	.931	.054	9.464E-02	10.566	9.464E-02
3	Populatia ocupata in 2000 (mii persoane)	.200[b]	.586	.590	.281	.457	2.187	.457
	PIB intern brut regional pe locuitor in anul 2000 (lei)	-.141[b]	-.604	.578	-.289	.975	1.025	.975

a. Predictors in the Model: (Constant), PIB intern brut regional pe locuitor in anul 2000 (lei), investitiile in 2000

b. Predictors in the Model: (Constant), investitiile in 2000

c. Dependent Variable: Castigul salarial nominal net in anul 2000

Beta in este coeficientul de regresie care ar rezulta dacă în pasul următor s-ar păstra în model variabila exclusă.

Statistica test t și *valoarea Sig.* sunt folosite pentru testarea ipotezei de nul cu privire la coeficienții de regresie, adică a ipotezei că între variabila dependentă și variabila independentă nu există o legătură semnificativă.

În exemplul considerat, se constată valori *Sig.* foarte mari (comparativ cu 0.05), ceea ce nu ne permite să respingem ipoteza de nul, a inexistenței unei legături semnificative între variabila dependentă – câștigul salarial – și

variabilele independente – populaţia ocupată şi PIB regional pe locuitor, la nivelul anului 2000, în România.

Se observă, de asemenea, valori mici pentru toleranţă şi valori mari pentru *VIF*, ceea ce denotă existenţa multicoliniarităţii care determină o varianţă mare a coeficientului de regresie, şi, ca urmare, o instabilitate a estimaţiei.

Respectarea ipotezelor cerute de analiza de regresie (erorile sunt distribuite normal, cu media zero; erorile au varianţă constantă; erorile sunt independente unele de altele) poate fi verificată grafic folosind diagramele *P-P Plot* şi *Scatterplot*. Figurile 9.12 şi 9.13 arată că sunt respectate aceste ipoteze.

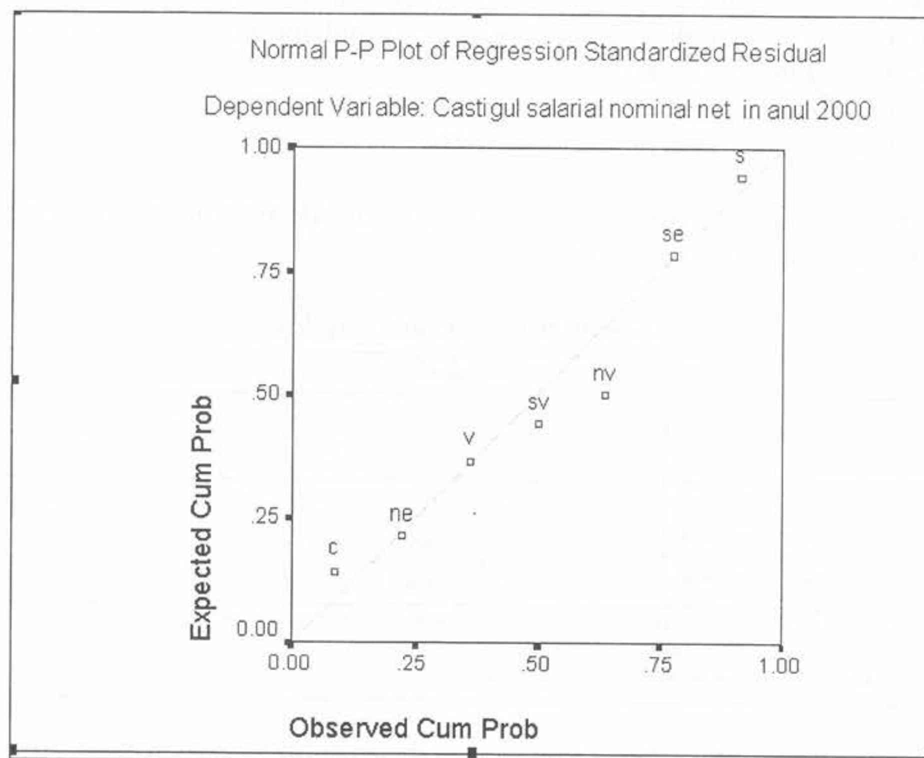


Figura 9.12 *Diagrama* Normal P-P Plot

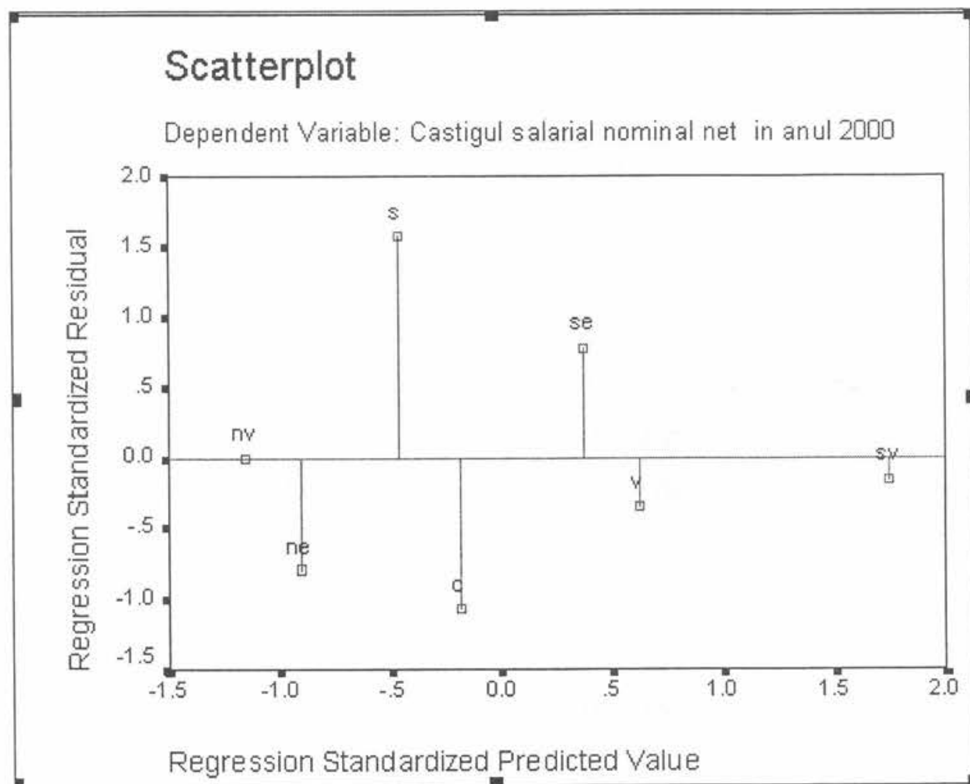


Figura 9.13 *Diagrama* Scatterplot

Bibliografie

1. Agresti, A., *Categorical Data Analysis*, John Wiley & Sons, New York, 1990.
2. Anderson, D.R.; Sweeney, D.J.; Williams, A.T, *Statistiques pour l'économie et la gestion*, De Boeck Université, Paris, 2001.
3. Andrei, T., *Statistică şi Econometrie*, Editura Economică, Bucureşti, 2003.
4. Bărbat, A., *Teoria statisticii sociale,* E.D.P., Bucureşti, 1972.
5. Begu, L.S., *Statistică şi Software statistic*, Editura Clauet, Bucureşti, 1999.
6. Berdot, J.P., *Économétrie*, Université de Poitiers, CNED, 2001.
7. Berenson, M.; Levine, D.; Krehbiel, T., *Basic Business Statistics*, Pearson Prentice Hall, New York, 2004.
8. Biji, M.; Biji, E.; Lilea, E.; Anghelache, C., *Tratat de statistică*, Editura Economică, Bucureşti, 2002.
9. Bourbonnais, R., *Économétrie,* ediţia a III-a, Dunod, Paris, 2003.
10. Bruhn, M., *Marketing*, Editura Economică, Bucureşti, 1999.
11. Dagnelie, P., *Statistique Théorique et appliquée*, Tome 2, *Inference Statistique à une et à deux dimensions*, De Boeck Université, Paris, 1998.
12. Dodge, Y., *Statistique. Dictionaire encyclopédique*, Dunod, Paris, 1993.
13. Field, A., *Discovering Statistics Using SPSS for Windows*, SAGE Publications, Londra, 2000.
14. Galton, Fr., *Natural Inheritance*, Macmillan, Londra, 1889.
15. Georgescu-Roegen, N., *Metoda statistică. Elemente de statistică matematică*, ediţia a II-a, Editura Expert, Bucureşti, 1998.
16. Grama, A.; Fotache, M.; Ţugui, A.; Dumitriu, F., *Medii de programare. Metode şi instrumente de dezvoltare a aplicaţiilor economice*, Editura Sedcom Libris, Iaşi, 2002.

17. Gujarati, D., *Basic Econometrics*, ediţia a III-a, McGraw-Hill, Inc., New York, 1995.
18. Jaba, E., *Statistica*, ediţia a III-a, Editura Economică, Bucureşti, 2002.
19. Kanji, G., *100 Statistical Tests*, SAGE Publication, Londra, 1993.
20. Korka, M.; Begu, L.S.; Tuşa, E., *Bazele statisticii pentru economişti*, Editura Tribuna Economică, Bucureşti, 2003.
21. Smith, G.M.; *Ghid simplificat de statistică pentru psihologie şi pedagogie*, E.D.P., Bucureşti, 1971.
22. Mitruţ, C.; Voineagu, V.; Isaic-Maniu, Al., *Statistica*, Editura Universitaria, Bucureşti, 2003.
23. Noica, C., *Jurnal filozofic*, Editura Humanitas, Bucureşti, 1990.
24. Oprea, D., *Analiza şi proiectarea sistemelor informaţionale economice*, Editura Polirom, Iaşi, 1999.
25. Piatier, A., *Statistique descriptive et initiation a l'analyse*, Themis, Paris, 1962.
26. Sora, V.; Hristache, I.; Mihăescu, C., *Demografie şi statistică socială*, Editura Economică, Bucureşti, 1996.
27. Tacu, Al.P.; Vancea, R., *Inteligenţa artificială*, Editura Economică, Bucureşti, 1998.
28. *** Anuarul Statistic al României, 2002, INS, Bucureşti, 2003.

Surse web

1. http://www.spss.com/
2. http://www.totalconsult.ro/spss.html
3. http://www.spss.ro/
4. http://www.csubak.edu/ssric/Modules/SPSS/SPSFirst.htm
5. http://s9000.furman.edu/mellonj/spss1.htm